CONTRIBUTIONS TO THE STATISTICAL AND COMPUTATIONAL MODELING OF DNA TRANSCRIPTION REGULATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

André Luís Martins

August 2014

CONTRIBUTIONS TO THE STATISTICAL AND COMPUTATIONAL MODELING OF DNA TRANSCRIPTION REGULATION

André Luís Martins, Ph.D.

Cornell University 2014

Transcription is a fundamental and tightly regulated process in living cells and a key step in the expression of the information contained in DNA. A wide variety of experimental assays have been developed that enable genome-wide analysis of the features of transcription and transcription regulation. We present statistical analysis combining both large existing datasets and new experimental assays to explore three aspects of transcription regulation: (i) determinants of transcription factor binding intensity, (ii) characterization of transcription initiation regions at both promoters and enhancers and (iii) unsupervised identification of transcription units.

Transcription factor binding intensity is affected by both DNA sequence and local chromatin landscape. We aimed to disentangle these influences by combining PB-seq (a new experimental approach developed by Michael Guertin) with existing modENCODE data in the study of *Drosophila* Heat Shock Factor (HSF). PB-seq enabled the estimation of the genome-wide binding energy landscape in the absence of chromatin. It further allowed the development of a statistical model to predict the departure of in-vivo binding intensities (from ChIP-seq) from the naked chromatin binding intensities (from PB-seq), based on covariates describing the local pre heat shock chromatin environment. We found that DNase I hypersensitivity and tetra-acetylation of H4 were the most influential covariates. Furthermore DNase I hypersensitivity could also be largely

recapitulated from the remaining covariates. Lastly, PB-seq data was applied to develop an unbiased model of HSF binding sequences, which revealed distinct biophysical properties of the HSF/HSE interaction and a previously unrecognized substructure within the HSE.

Transcription initiation regions at promoters and enhancers have conventionally been treated separately, although they share many features in mammals. We examined all transcription initiation sites, for both stable and unstable transcripts, using GRO-cap (a new experimental assay developed by Leighton Core). Statistical modeling and analysis of this data, and its contrast with existing ENCODE datasets, reveal a common architecture of initiation at both promoters and enhancers. This common architecture features tightly spaced (110 bp) divergent initiation with similar frequencies of core-promoter sequence elements, highly-positioned flanking nucleosomes, and two modes of TF binding. Transcript elongation stability, a feature determined after transcription initiation, provides a more fundamental distinction between promoters and enhancers than the relative abundance of histone modifications and the presence of TFs or co-activators. These results support a unified model of transcription initiation at both promoters and enhancers.

Finally, we turn to the identification of transcription units from nascent RNA assays (GRO-seq and PRO-seq). Although existing annotations focus on stable RNA transcripts (cleavage and poly-Adenylation point), transcription extends beyond the cleavage site. As such, the transcription process can potentially influence surrounding regions. We improve on previous work on the detection of transcription units by obtaining an unsupervised method that does not depend on RNA product annotations. We use these results to examine post poly-Adenylation extension and cross-strand RNA polymerase collision effects.

BIOGRAPHICAL SKETCH

André was born and raised in Lisbon, Portugal. There he attended university at Instituto Superior Técnico where he graduated in 2004, in Computer Science and Engineering (5 year Lic.), and later, in 2006, he obtained a MSc. in Computer Science with a focus on Machine Learning. During that time he started his research at INESC-ID and joined the bioinformatics group there (KDBIO). This experience, together with the encouragement of his advisor Prof. Arlindo Oliveira, led him to move to the US in 2007 to pursue a Ph.D. in Computational Biology at Cornell University. There he joined Adam Siepel's lab where he conducted his research into transcription regulation. This research benefited from the tight collaboration with students from John Lis' lab. Outside the lab, André enjoys training in the Japanese martial art Aikido at the Cornell Aikido Club.

To Susana

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents for their constant and unconditional support and encouragement during this work and throughout my life.

This thesis would not have been possible without the help, support and patience of my advisor, Prof. Adam Siepel. His good advice and expertise have been invaluable on both an academic and a personal level, for which I am extremely grateful. I owe a debt of gratitude to Prof. John Lis, who significantly helped me with his advice and expertise and by allowing my working together with the transcription regulation subgroup of his lab, thus stimulating great collaborations that had a positive impact on this thesis. I also wish to acknowledge the other member of my committee, Prof. Giles Hooker, for his support and input provided on several occasions and during my A-exam.

I would like to thank all the kind people around me who contributed in some way to the achievement of this thesis. Chapters 2 and 3 in this thesis are a direct result of a productive collaboration with Michael Guertin and Leighton Core, respectively. I also benefited from invaluable support and many interesting discussions with Charles Danko and all members of the transcription regulation subgroup of Prof. Lis lab. Jason Mezey and Sean Myles helped me during my brief foray into the experimental biology as I rotated through Mezeys lab. I also wish to remember Prof. Arlindo Oliveira and my colleagues of the KDBIO group at IST/INESC-ID, U. of Lisbon, where I took my first steps in research in this field.

I would like to thank Susana for her love and continued support and encouragement.

Finally, I gratefully acknowledge the support of this research from Fundação para a Ciência e a Tecnologia (Portugal) (SFRH/BD/30980/2006), Fulbright Portugal, the Cornell University Provost, and research projects funded by the National Institute of Health and National Science Foundation and lead by Profs Adam Siepel and John Lis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Proteins, nucleic acids and polysaccharides are among the essential macromolecules contributing to the fundamental functions in living organisms. Proteins are the main actors in cell function and participate in virtually every process within the cell. Some proteins are enzymes that catalyze biochemical reactions and are vital to metabolism; other proteins have mechanical (structural) functions in the cytoskeleton, or participate in transport processes, cell signaling, gene regulation, etc.

The production of proteins within the organisms and the regulation of their functions are thus critical aspects of life. Francis Crick, in 1956, named Central Dogma the pathway relating DNA, RNA and proteins, where the main processes identified are DNA replication, DNA to RNA transcription and RNA to protein translation.

Not all DNA is subject to transcription; in fact, most transcription is focused on relatively small sequences, called genes, which encode the information needed to assemble proteins. Nevertheless, other regions of the genome are also transcribed to a lesser extent and are thought to play a role in primary protein-coding gene regulation. In particular, unstable (rapidly degraded) transcripts are observed upstream of protein-coding genes (sometimes referred to as upstream anti-sense non-coding RNAs or uaRNAs [109][27]) and at distal enhancer transcription factor[1] binding sites (sometimes referred to as enhancer

[1] A transcription factor is a protein, or protein complex, that binds DNA, either directly or indirectly, and influences the amount of transcription either locally or at a distal target. Distal enhancers where the regulatory regions bound by these factors that have positive effect on transcription initiation of their target genes.

RNAs or eRNAs [78]). Furthermore, non-coding transcripts, such as long interspersed non-coding RNAs (lincRNAs [102]), have been a recent focus of attention as new methods have enabled their detection. These RNAs can be found either by themselves or coupled with protein-coding genes (in place of uaRNAs) or with eRNAs (short non-coding RNAs produced at enhancer regions), at enhancer binding sites. For all of these, RNA Polymerase II (Pol II) is the main molecular complex responsible for transcription, following a similar, highly regulated, process.

An initial level of transcription regulation occurs through DNA packing around protein complexes called nucleosomes (the DNA-nucleosome complex is known as chromatin, although this term is also used to include other DNA-binding proteins). Formed by two copies of each of the core histones H2A, H2B, H3, and H4 [92], the nucleosome resembles two parallel disks around which DNA loops twice, totaling about 146 bp. Various features of these histones, including their placement along the DNA sequence, composition (there are variants of the core histones), and modifications of the free "tails" of the histones, result in varying degrees of DNA packing. Thus, through modulation of DNA accessibility, parts of the genome are made available or unavailable to transcription initiation and regulatory factor binding. DNA accessibility is thought to be one of the main forms through which cell differentiation propagates regulatory state during development of multi-cellular organisms [130].

In the presence of suitably accessible DNA regions, Pol II can then be recruited by transcription factors at the core promoter region and proceed through its the transcription activity cycle [45] (see Figure 1.1). The cycle starts by the binding of the core transcription factors, which help recruit Pol II, forming with

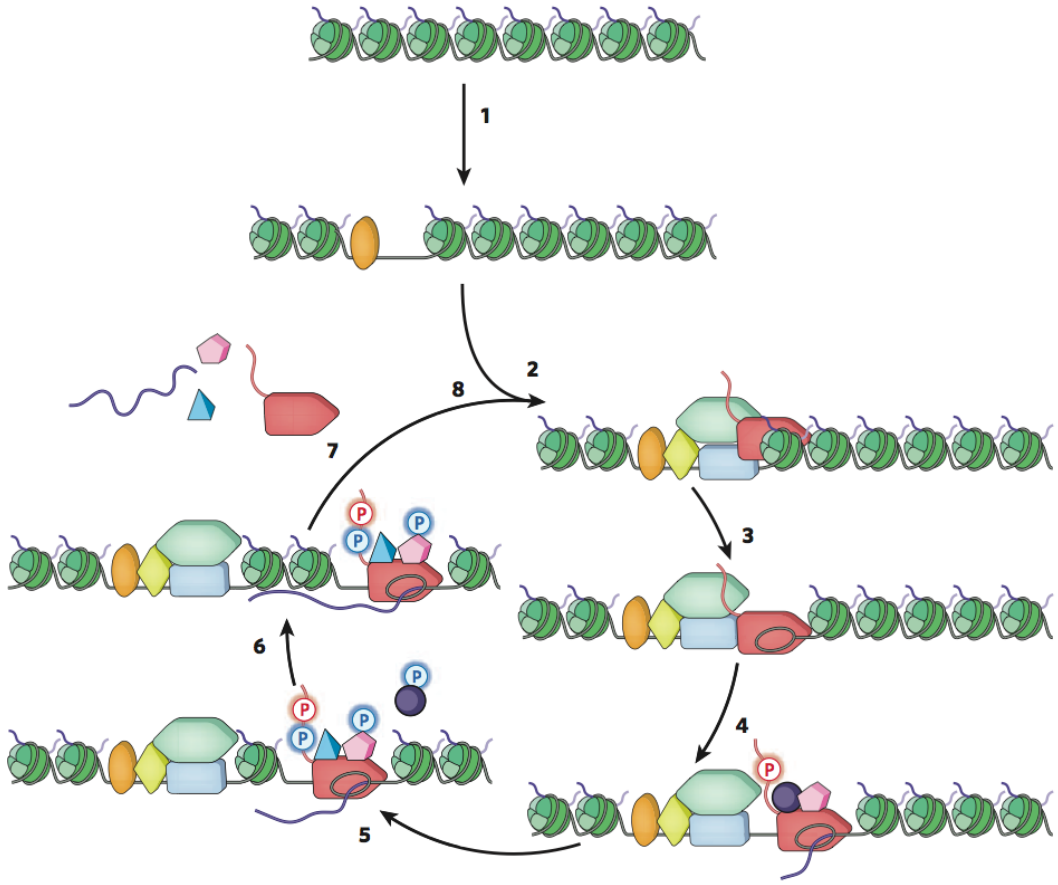

Figure 1.1: Transcription cycle: 1) Chromatin opening; 2) PIC formation; 3) Initiation; 4) promoter clearance; 5) escape from arrest; 6) productive elongation; 7) termination. Green balls represent nucleosomes; red rocket represents polymerase; Ps represent phosphorylations; purple strand represents RNA; other shapes represent transcription factors. [Adapted from Fuda et al., Nature 461 (10), 2009]

it the pre-initiation complex (PIC). The formation of the PIC, and whether or not it results in actual productive transcript elongation, often depends on which protein transcription factors (TFs) are associated with both the promoter region and more distal enhancer and repressor regions. These transcription factors interact directly or indirectly with the PIC and are a component of the transcription regulation. The next step in the cycle, promoter-proximal pausing, shortly after elongation initiation, has been observed on a subset of genes, in several

organisms. Pol II pausing and accumulation for extended periods is observed around 50 bp from the transcription start site (TSS). It constitutes a regulatory step that allows rapid "initiation" of gene transcription following an external stimulus mediated by a transcription factor, by bypassing the relatively slower process of assembling the PIC and transitioning into transcript elongation [1]. This is followed by an elongation phase where co-transcriptional processes like intron splicing occur [113]. This phase is completed with the recognition of sequence elements that mark the end of the mRNA transcript (or early transcription termination for unstable transcript products) and the cleavage addition of a polyA tail (long sequence of adenine nucleotides) [83]. The transcription activity cycle however does not end here, as Pol II continues to transcribe beyond this point, producing an independent transcript, until actual transcription termination. One of the proposed processes for this final termination is a collision between the transcript degradation machinery and Pol II, as it concurrently consumes the nascent uncapped transcript [115]. Pol II is then free to restart the cycle again. All steps in the cycle are subject to regulation through direct and indirect interactions with proteins binding the DNA, the nascent RNA, nucleosomes or the Pol II complex itself.

As mentioned, chromatin accessibility is a main factor in transcription regulation, by either making transcription regions accessible to the Pol II initiation complex or making distal regulatory regions accessible for transcription factor binding. The latter is dependent on physical proximity of distal regions and is subject to chromatin-chromatin bridging by factors like CTCF and cohesin [98]. Naturally, chromatin structure and composition has been a major focus of study and a driver for important technology development.

In recent years, the development of short read sequencing technology coupled with Chromatin Imunoprecipitation (ChIP-seq) [69] enabled measurement of DNA-protein associations on a level not previously possible. It has allowed measurements of the distribution across the genome of histone variants, post-translational modifications of histones, bound transcription factors (TF), transcriptionally engaged Pol II, and other important components of transcription regulation.

In addition, experimental assays have been developed to capture the RNA products of transcription. These assays can be broadly divided into those based on stable accumulating mRNA molecules (messenger RNA, the end product of gene transcription) and those based on mapping nascent RNA (newly synthesized RNA still attached to elongating Pol II). The first group includes RNA-seq, which is used to map the transcribed regions (exons in the case of protein-coding genes), and CAGE-seq, which is used to precisely map the start sites of these stable mRNAs. The second group, which has developed more recently, includes the global nuclear run-on assay (GRO-seq; [27]), which has enabled the quantification of RNA synthesis distribution across the genome by engaged polymerase molecules (among them, Pol II). GRO-seq helps map both the DNA that is transcribed as part of mRNAs, the spliced out portions (introns) and, most importantly, the rapidly degraded transcripts (uaRNAs and eRNAs). This assay, and its higher resolution descendent PRO-seq (precision nuclear run-on and sequencing; [81]), maps the position of engaged Pol II and relative amounts present across the genome. These assays thus enable the inference of the presence of obstacles to transcription and their relative strength, and are informative about cross-strand collisions between engaged Pol II.

The development of genome-wide assays and sequencing technologies has enabled large-scale analysis of many genomes. The goals in genomics are now, not only to obtain the full genome sequence for many organisms and annotate their genes, but also to map TF binding sites, chromatin state and accessibility, transcribed regions, etc., across cell types, developmental stages and experimental conditions — that is, to get a full picture of the genome and associated proteins, to enable a more complete understanding of genome structure and (transcriptional) regulation.

Leading this effort are two key projects, ENCODE and modENCODE. The aim of these two projects is to gather the results of the efforts of researchers around the world into the understanding of the human (through the ENCODE project) and the major animal model organisms, *D. melanogaster* and *C. elegans* (through the modENCODE project). The efforts of these large consortia provide a wealth of information that can be used to further our models of transcription and gene regulation.

The work reported in this thesis aims to leverage advances in sequence-based assays and the large datasets available to explore transcription regulation. Chapter 2 is focused upon the differences between the TF binding affinity to naked DNA versus in-vivo chromatin modulated binding affinity, through a detailed study of Heat Shock Factor (HSF) binding in *Drosophila*. Chapter 3 focuses on transcription initiation and uses a variation on a nascent RNA based assay to characterize divergent transcription initiation site structure, in human cells, for all Pol II derived transcripts independent of stability. Finally, chapter 4 focuses on unsupervised detection of transcription units (TU) from nascent RNA and analysis of TU profiles with a focus on post-polyA extension. The

following sections provide a brief background on technical topics needed to understand the methods used throughout this work.

1.1 Transcription Factor Binding Models

Research into transcription factor binding is a vast field from both the experimental and computational perspectives. Ultimately, the answers to questions such as which transcription factors are bound, where are they bound across the genome and how strongly are they bound at each locations, are necessary for a comprehensive understanding of transcription regulation. As such, these questions have been approached through both experimental and computational methods. This section will give a very brief description of the relevant aspects of transcription factor binding and the commonly used approaches to elicit experimental binding data and to model the sequence affinity of transcription factors.

Upon activation and entering the cell nucleus, a DNA binding transcription factor (TF) protein complex exists in a chemical equilibrium with accessible DNA[2] (the substrate S):

$$\mathrm{TF} + \mathrm{S} \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} \mathrm{TF} : \mathrm{S}$$

In other words, part of the time the TF will be bound to a particular DNA sequence (DNA sequence element, or substrate S; bound state denoted as $TF : S$) and otherwise it will be free (as denoted by the '+' sign). The relative time that the TF is bound will depend on factors such as the concentration of the factor ([TF]), the concentration of accessible DNA ([S]) accounting for relative time that the DNA is accessible (their may be other proteins competing for binding, such

[2]Some transcription factors do not bind directly to DNA, but rather to some other protein complex. Replacing DNA with the target protein will yield a similar description of the process.

as nucleosomes) and the binding affinity of the TF and the underlying DNA sequence. This quantity, the dissociation constant (K_d), is defined as:

$$K_d = \frac{k_{off}}{k_{on}} = \frac{[TF][S]}{[TF:S]}$$

The most commonly used experimental assay to detect binding in in-vivo conditions is ChIP-seq [69]. Briefly, in ChIP-seq, TF-DNA binding is chemically frozen in place by a process called cross-linking and, through immunoprecipitation, the segments of DNA to which the TF is bound are separated out and subjected to short-read sequencing. Sequenced reads are mapped back to the genome and preferentially align ("pile-up") around the binding sites. Software, such as MACS [149], has been developed to identify these pile-ups of aligned reads, and therefore detect "peak" regions on the order of a two to four hundred base pairs that putatively contain as binding sites. Note that most TFs recognize a very short DNA sequence, typically between 5 and 15 bp, so peak identification is just an initial step in the analysis. However, peak identification is not without its difficulties. As in-vivo TFs are interacting with their regulatory targets, cross-linking will to some degree indirectly select for these distal regulatory targets, leading to false-positives. Furthermore, it is not uncommon to have multiple binding sites in close proximity (a few hundred base pairs or less), which leads to partially overlapping ChIP-seq peaks, usually detected as a single broad peak. Finally, ChIP-seq, as with many modern assays, is conducted not for an individual cell but for a population of cells, so the read profile reflects a mixture of binding events across these cells. Ongoing research is being conducted to improve resolution, through methods like ChIP-exo [114], and the use of factor agnostic information such as footprinting methods based on high density DNAse-seq [63][105].

TF sequence affinity is typically modeled through a position weight matrix (PWM), which describes the relative frequency of each nucleotide per binding position, under a simplifying position independence assumption, and can be related to binding energy [14][132]. PWMs are either obtained experimentally by successive selection steps over a pool of small random DNA fragments (SELEX and related methods; [131][118]) or computationally, by searching for enriched patterns from a set of experimentally obtained DNA sequences, for example ChIP-seq peaks, given some background model (MEME [6] and similar methods). The former approaches tend result in PWMs that more closely reflect the TF sequence preferences, while the latter approach tends to reflect the genomic distribution of binding sequences for that factor. These are not likely the same, as tuning sequence affinity is a possible way through which natural selection can fine tune transcript regulation. Several attempts have been made to build models that allow for richer position dependencies without overfitting [9][95], but they are not currently in widespread use. Furthermore, TFs are sometimes composed of multiple binding domains, or work as a complex of two or more copies, with potentially variable spacing in-between [95]. Sometimes the problem can be made easier through the addition of a sequence set where binding does not occur or is expected to be depleated [137]. Overall, given a PWM (or set of PWMs) the next step is to search for binding sites, i.e., sites along a set of potential sequences, where sequence composition better matches the PWM than the background model.

TF binding site identification is a hard problem. Typically one is not interested in all potential sites across the genome, but rather the much smaller subset that is actually bound in a given condition. As TFs are not the only proteins compeeting for DNA binding, it is not necessarily the case that the best binding site

in a local region is actually the one bound by the factor of interest, although this is a reasonable heuristic in many cases. Common approaches include focusing TFBS identification on ChIP-seq peaks, possibly including a peak deconvolution step, or baring the availability of ChIP-seq data for the factors in question, reusing DNase-I HS peak information (e.g Centipede method; [106]). Finally, TF binding sites (TFBS) can be part of larger regulatory modules (see [150] for an example), which can potentially be used to improve detection.

1.2 Linear Regression

Linear regression models, and generalized linear regression models such as logistic regression models, are often used as a first step towards understanding how a quantity of interest Y (response) relates to some observed set of features or covariates X. The simple model structure lends itself to ease of parameter inference and interpretation, especially with lower dimensional inputs. In general terms, the regression function $r(x)$ can be defined as:

$$r(x) = E[Y|X = x] = \int y f(y|x) dx$$

where $f(y|x)$ is the conditional probability of the response given the covariates. Given a number of observations (Y_i, X_i), the goal is to estimate the regression function. Linear regression models take the form:

$$Y = X\beta + \epsilon \quad \text{or} \quad E[Y] = X\beta$$

where Y and ϵ are vectors of length N, β is a vector of length k, and X is a $N \times k$ matrix. Also, $E[\epsilon_i|X_{i\bullet}] = 0$ and $Var[\epsilon_i|X_{i\bullet}] = \sigma_i$. Generalized linear models extend this definition to:

$$E[Y] = \mu = g^{-1}(X\beta)$$

where $g(\mu)$ is named the link function. In particular, for logistic regression, the link function is defined as:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) \quad \text{or} \quad E[Y] = \frac{1}{1+\exp(-X\beta)}$$

The parameter vector β can be inferred through least squares, for linear models with the error term following a Normal distribution, or using the reweighted least squares algorithm for logistic regression. In either case, we obtain a set of estimated parameters (regression coefficients). Direct interpretation of the coefficient values is useful, but still requires some effort. In particular, it is necessary to normalize the input data such that all input covariates are on the same scale, to ensure that the coefficients are comparable. It is also possible to test the hypothesis that each coefficient is zero, thus interpreting the p-values to indicate that the corresponding covariates are useful. One alternative approach is to estimate the fraction of explained variance (R^2) attributable to each covariate [51][135] (a similar method has been developed for logistic regression [70]).

In genomic studies, however, there are usually a very large number of covariates and some interest in modeling interactions. Feature interactions are generally problematic to handle due to the exponential growth of the number of interaction tuples as the tuple size grows. This combined with the assumption that many of these features/combinations are not relevant, leads to the use of sparse regression, where a penalty is imposed on regression coefficients that push them towards zero. Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani [134], is a popular penalty for feature selection in linear regression models. One way to express this penalty is to add a term to

the conditional log-likelihood form of the regression:

$$\begin{aligned} l(\beta) &= \ln P(Y|X) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{y_i\left(\beta_0+\sum_{j=1}^{k}\beta_j X_{ij}\right)-\ln\left(y_i\left(\beta_0+\sum_{j=1}^{k}\beta_j X_{ij}\right)\right)\right\}-\lambda P(\beta) \end{aligned}$$

where $P(\beta)$ is the penalty of the form: $P(\beta) = \sum_{i=1}^{k} |\beta_i|$ and λ is the penalty weight (usually tuned via cross-validation). Over the years, LASSO has been extended to several classes of generalized linear models, including logistic regression [42] and more elaborate models that extend linear regression to incorporate more advanced models of feature interaction. One such method, *rule ensembles* [43], used in Chapter 2, combines small conjunctive rules (decision trees) with linear regression and applies a LASSO penalty to keep the number of rules under control. As such, it strikes a compromise between modeling interactions in an interpretable way and exploring the set of possible interactions. Similar to regular linear modes, it is also possible to define a measure of relative importance for covariates in the context of rule ensembles that takes into account not only the coefficients of each rule, but also in what fraction of the input data do the rules activate. This is discussed in more detailed in [43].

Overall, linear regression models and their extensions provide flexible tools to express the relationship between properties of interest, with modest computational and parametric complexity, making them widely applicable not only in the interpretation of biological data but also in many other fields.

1.3 Hidden Markov Models

A hidden Markov model (HMM) defines a probability distribution over a, possibly infinite, set of finite sequences of values. Here a value is anything over which a probability distribution can be defined. A Markov model is a set of random variables, one per position along the sequence, that form a conditional dependency chain between consecutive positions. That is, the probability of the random variable (Z_i) is defined conditional on the value of the random variable at the previous position (Z_{i-1})[3] : $P(Z_i|Z_{i-1} = v)$, together with an initial state probability distribution $P(Z_i)$ (see Figure 1.2A). Given this structure, the chain can be interpreted as a random walk though the state space (the domain of Z_i) as time passes on (each position being a time point); in other words, a generative process where each successive state (Z_i value) is obtained from the previous state (Z_{i-1} value) according to the probability $P(Z_i|Z_{i-1})$. In a HMM, at each position there is an additional random variable X_i generated (emitted) based on the state, according to the distribution $P(Z_i|Z_{i-1} = v)$. The sequence of X_i is referred to as the observations, or emissions, (typically known values) and the sequence of Z_i is referred to as the hidden state path (typically unknown). One way to interpret the model is to consider the observations as a noisy readout of the true hidden state sequence. HMMs are a very useful framework for sequence segmentation and find applications in areas like speech recognition, DNA sequence analysis, among others [110][34].

The hidden state space can be either continuous or discrete, depending on the application. Similarly, the above definition can be extended from unit time steps (discrete time) to varying continuous steps (continuous time). In the later

[3]This is a first-order Markov model. Higher order models are also possible, with Z_i depending on the "k" previous positions, i.e., $P(Z_i|Z_{i-1}, \ldots, Z_{i-k})$.

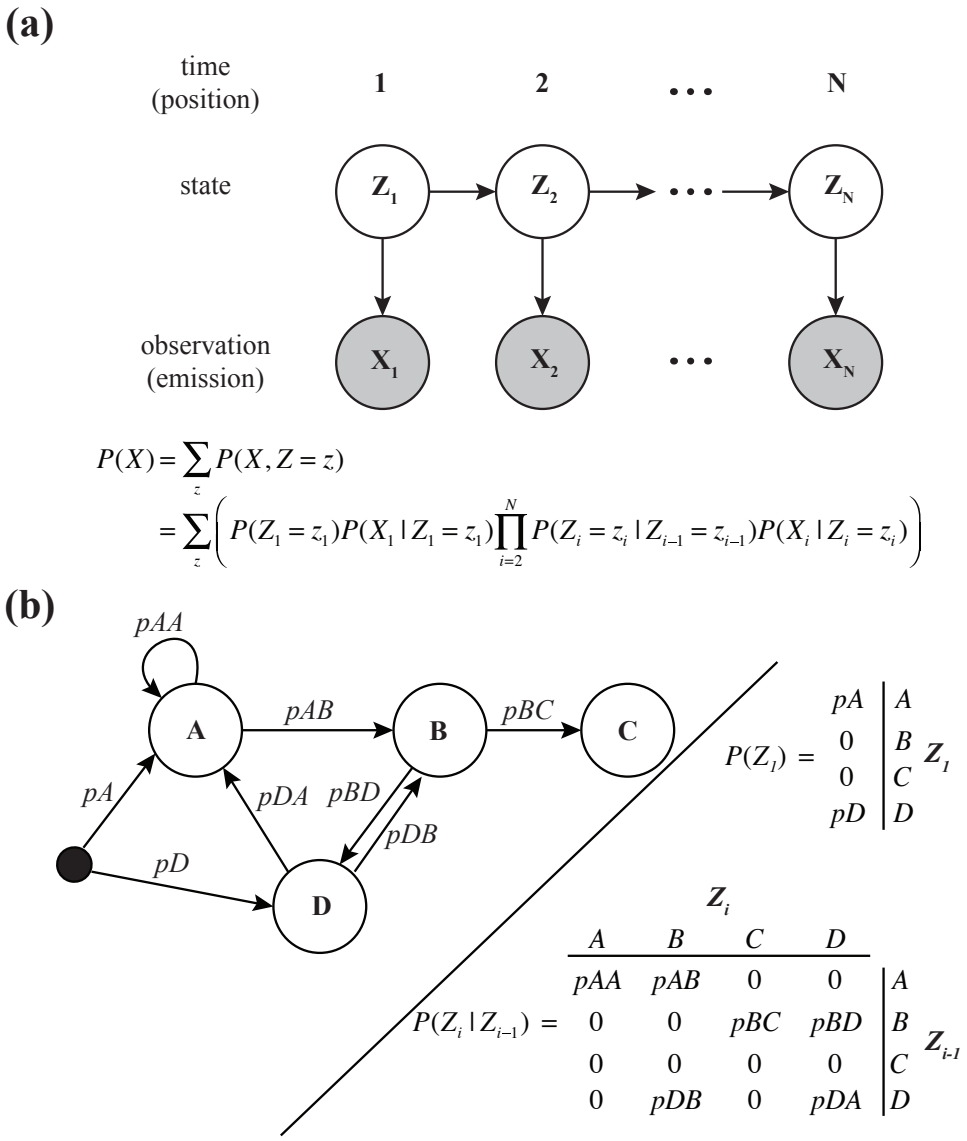

Figure 1.2: Hidden Markov Model random variable and state graphs. (a) HMM random variable graph for a sequence of length N: states $Z_1 \ldots Z_N$, emissions $X_1 \ldots X_N$. Observed nodes shaded in light grey. Also shows the implied sequence likelihood function definition $P(X)$. (b) Example finite automata for a finite discrete state space. Valid transitions (non-zero transition probability) indicated by arrows (i.e., an arrow from A to B implies $P(Z_i = B|Z_{i-1} = A) > 0$). The right side shows the corresponding state initial probabilities ($P(Z_i)$) and transition probabilities ($P(Z_i|Z_{i-1})$).

case, each state change is associated with an event time that drawn from an exponential distribution. In this work, we restrict ourselves to discrete time and (finite) discrete state spaces. With a discrete space HMM, the state space can be represented by a (regular) automaton (see Figure 1.2B), with transition from state A to state B corresponding to a non-zero probability $P(Z_i = A|Z_{i-1} = B)$. If the definition of these transition probabilities is identical across the sequence the HMM is said to be homogenous, otherwise it is called non-homogeneous.

Typically, the hidden states are unknown and we have one or more sequences of observations (emissions) with the goal of fitting the model to the data (parameter inference) and determining the hidden state path (parsing). Parameter inference is easy if the matching set of hidden states are known, as parameters can then be directly estimated. When the hidden state paths are unknown, the maximum likelihood parameter estimates for transition and emission probability distributions, given a set of observed sequences, are obtained via Expectation-Maximization (EM) algorithm [32]; a well-known instance of EM, when both the state space and the observation space are finite discrete sets, is the Baum-Welch algorithm [11][34]. After obtaining a set of parameter estimates, the HMM model can be used to parse or decode an observed sequence into a path through the hidden state space. Given a set of parameters and an observation sequence, the maximum likelihood path (Z) can be obtained via the Viterbi algorithm [139]. An alternative way to parse an observation sequence is to compute, at each position, the posterior probability $P(X_i|Z)$ and then chose the state with maximal posterior probability, known as posterior decoding. This does not necessarily yield a valid path through the state space. An example application of these algorithms to a simple two-state model can be seen in Figure 1.3.

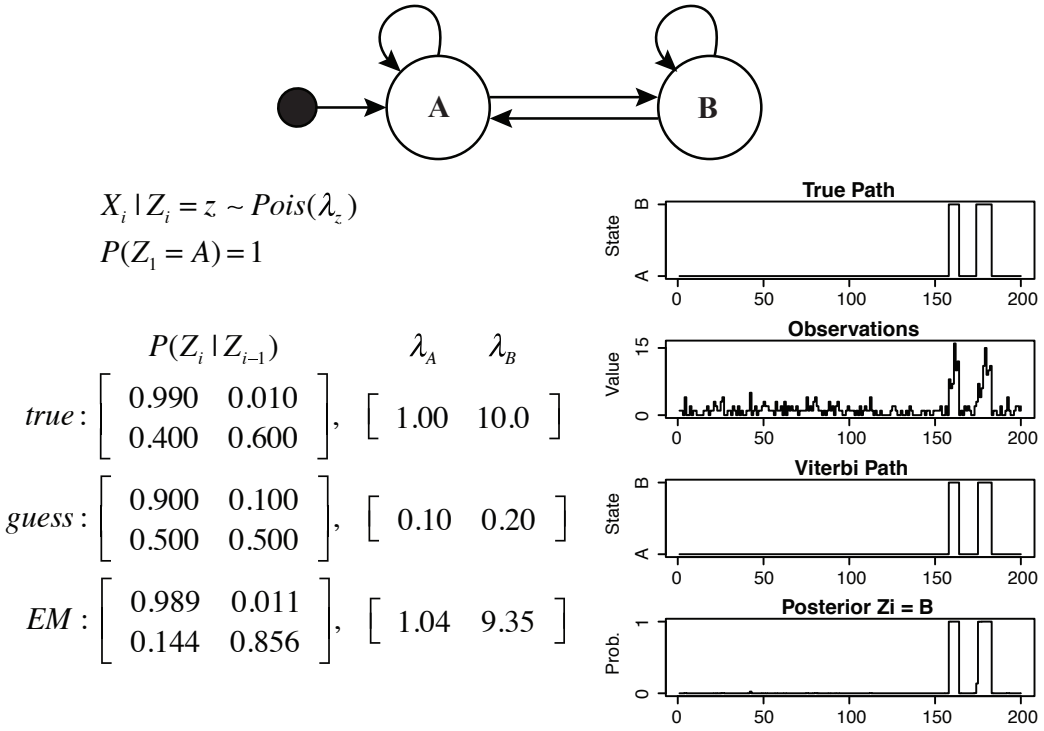

Figure 1.3: Example two-state HMM with Poisson emission distributions. Transition and emission parameters are shown on the left for the true HMM (used to generate state path and observations shown on the top two right-side plots), the initial guess values used for EM and the final EM estimates. Additionally, the bottom two right-side plots show the viterbi path and the posterior decoding using the estimated EM parameters.

In practice, HMM models have a number of shortcomings. When used for segmentation, it is not always the case that the length distribution of the segments (particular sequences of hidden states of interest) can be well approximated by regular grammars (combinations of geometric distributions per state). Parameter inference can be challenging. Although EM is a very effective technique, it is not guaranteed to find the global maximum over the entire parameter space, so parameter initialization plays an important role. In practice, this comes into play due to the interplay between transition parameters and emission parameters, namely, a bad initialization or model structure, may lead transitions into a particular state to have zero probability. For that reason, some-

times it is necessary to fix parts of the model parameter set using some a priori knowledge and not estimated via EM. This section introduced the basic concepts and terminology to help an unfamiliar reader to follow the uses of HMMs in the following chapters. It is not, by any means, a summary of the vast field of HMM applications.

CHAPTER 2

ACCURATE PREDICTION OF INDUCIBLE TRANSCRIPTION FACTOR BINDING INTENSITIES IN VIVO[1]

2.1 Introduction

Binding of transcription factors (TFs) to DNA elements is necessary to establish and maintain functional changes in gene expression levels. The mechanism by which these factors seek out and bind to their cognate motif elements remains an area of active investigation (reviewed in [39]). TFs are present at cellular concentrations that allow binding to sites that are degenerate from the consensus sequences, and genomes of eukaryotes are littered with potential degenerate binding sites; however, only a small fraction of potential binding sites are recognized in vivo. Moreover, TF binding sites vary dependent upon cell type and cellular conditions. In vivo, TF binding is potentially dependent upon motif accessibility and the surrounding chromatin landscape. Therefore, determining a comprehensive set of potential genomic binding sites and quantifying the joint effects of DNA sequence and chromatin landscape upon binding intensity remains a challenge.

Experimental approaches to characterize TF binding sites include assays such as ChIP-seq, protein binding microarrays (PBM) [15], iterative rounds of protein-DNA binding and selection with a complex oligonucleotide library [89], or extrapolation from DNase I hypersensitivity regions [63]. However, perhaps

[1]The content of this chapter is the result of a joint work with Michael Guertin, published in PLoS Genetics 8(3): e1002610, March 2012. Michael Guertin developed and conducted the experimental assays. I developed the statistical analysis. Writing and interpretation were a joint effort.

the most direct way to determine all potential TF binding sites within a genome is to incubate purified TF and naked sheared genomic DNA in vitro, and then specifically quantify the TF-bound DNA [90]. This in vitro method allows binding sites to be interrogated in their native sequence context without the confounding effects of chromatin and cooperation between chromatin-bound factors.

It is challenging to predict in vivo TF binding accurately even when all potential in vitro binding sites have been characterized, because the chromatin landscape dramatically influences binding and it changes dynamically with development and with alterations in cellular nutrition and environment [53], [86]. Recent TF binding site modeling efforts have considered genomic nucleosome occupancy or DNase I hypersensitivity data to account for the effect chromatin has upon in vivo TF occupancy [73][99][106][18].

However, these models are limited in that they rely upon genomic accessibility data and TF binding data produced under the same conditions. To date there are no data sets that describe the full set of potential TF binding sites, the chromatin state data prior to binding, and occupied binding sites in vivo, in a single inducible system. Integration of these three data sets would allow one to decouple the effect TF binding has upon chromatin state from the effect preexisting chromatin state has upon induced TF binding.

The heat shock response of *Drosophila* is a model system extensively used to study the general functions of sequence specific activators and how they function to regulate transcription (reviewed in [54]). The master regulator of the heat shock genes, Heat Shock Factor (HSF), has a modest affinity for DNA under non-stress conditions [53][58][48], and upon stress, HSF homotrimerizes and in-

ducibly binds to a conserved consensus motif at over 400 sites in the *Drosophila* genome [53][48]. While over 95% of the HSF binding sites contain an underlying HSF sequence motif element (HSE), the vast majority of predicted genomic HSEs remain HSF-free following heat shock. Therefore, the chromatin landscape most likely plays a prominent role in determining binding of HSF.

Here, we describe an experimental technique, protein/DNA binding followed by high-throughput sequencing (PB-seq), to quantify the binding potential of all binding sites within a genome. We then develop a quantitative model that incorporates HSF PB-seq data, together with HSF ChIP-seq in *Drosophila* S2 cells [53] and S2 cell chromatin data, that accurately predicts observed in vivo HSF binding profiles. Moreover, our model allows us to quantify the relative importance of the chromatin features influencing HSF binding intensity. Finally, we develop a sequence model that uses HSF PB-seq data to characterizes the relationship between positions within the HSE and provide biophysical insight into the mechanisms by which HSF interacts with its cognate element.

2.2 Results

2.2.1 Quantification of the absolute binding affinity of all genomic *Drosophila* HSEs

We performed an in vitro binding experiment with purified HSF (Figure A.1) and naked, sheared genomic *Drosophila* DNA, to derive an accurate set of potential HSF binding sites in the *Drosophila* genome. HSF-bound DNA was specifi-

cally eluted and detected by high throughput sequencing (Methods). The HSF PB-seq experiment yielded 68% of the sequence tags within peaks. In contrast, typical ChIP-seq protocols are more inefficient and the majority of DNA (60% to >99%) sequenced is uninformative background DNA [104].

Peak calling revealed 3952 HSF-binding peaks ($p < 0.01$; 2848 peaks were common to both experimental replicates), which include 60% of the previously identified high-confidence HSF binding peaks in vivo [53]. The naïve expectation is that every in vivo HSF peak should have a corresponding in vitro peak, but it is not surprising to observe an incomplete overlap of in vivo by in vitro peaks, for various reasons. As will be discussed, binding sites detected in vivo but not in vitro tend to be more degenerate and have higher DNase I accessibility. Additionally, in vivo binding sites that are dependent upon cooperative interactions with pre-bound chromatin factors, long range DNA interactions, post-translational modifications of HSF [22], higher-order chromatin structure, or bridging protein interactions [49] will not be detected in the current form of PB-seq.

Underlying the in vitro binding peaks, we detected 3735 clusters of HSF binding site HSE sequences (2896 in peaks common to both replicates) at 20% HSE False Discovery Rate (FDR). We used clusters of co-occurring sites due to the uncertainty in HSE detection (see Methods). Furthermore, the majority, 3389 clusters (2586 in peaks common to both replicates) are not detectably bound in S2 cells in vivo. Figure 2.1 shows two examples of in vitro binding sites flanking the Cpr67B gene that are not bound in vivo. Moreover, the in vitro binding data quantifies differences in the in vitro and in vivo HSF binding intensity, such as the peaks within each of the promoters for Hsp23 and Hsp26 (Figure 2.1).

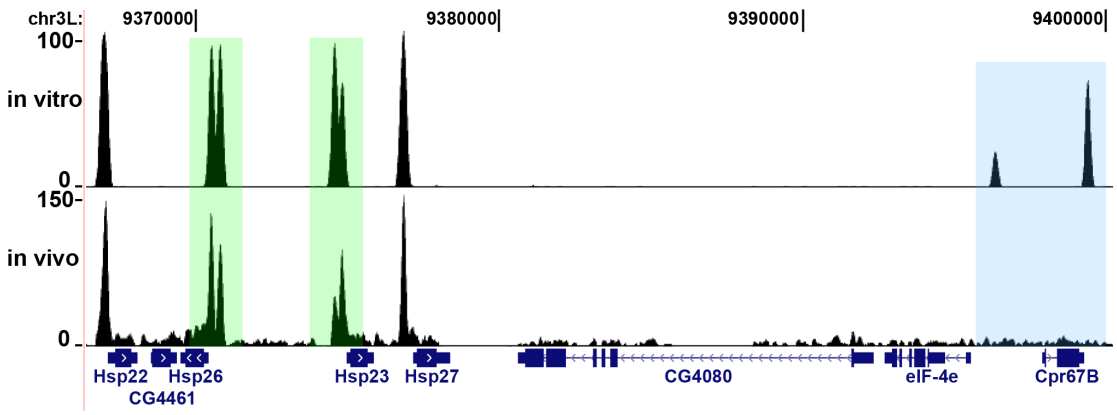

Figure 2.1: In vitro binding reveals potential HSF binding sites. The blue box highlights strong differences in the usage of potential binding sites in vivo at the Cpr67B locus, while the green boxes highlight differences in the magnitude of binding to major heat shock genes promoters, despite comparable in vitro binding affinities.

The PB-seq experiment allows for an estimate of the relative binding intensity of each HSE, based on the number of sequence tags associated with it. To compute the dissociation constant (Kd) values it is necessary to have estimates for both the fraction of bound and free HSE in the PB-seq experiment. Since the PB-seq data only provides information on the bound fraction, we needed to determine the absolute Kds for two HSEs that are found within the PB-seq data in order to provide enough information to estimate the free fraction (see Methods).

To generate the HSF/HSE Kd measurements, we performed electrophoretic mobility shift assays (EMSA). The EMSAs were performed with purified HSF and HSEs that are only modestly degenerate from the consensus. We found that HSF binds to the first HSE with ~ 42.6 pM interval: 36.8-49.4 pM; Figure 2.2A and Figure 2.2C) and the second HSE with ~ 224 pM affinity (95% confidence interval: 181-276 pM; Figure 2.2B and Figure 2.2D). The resulting two absolute Kd values enabled us to transform PB-seq read depths into absolute Kd values

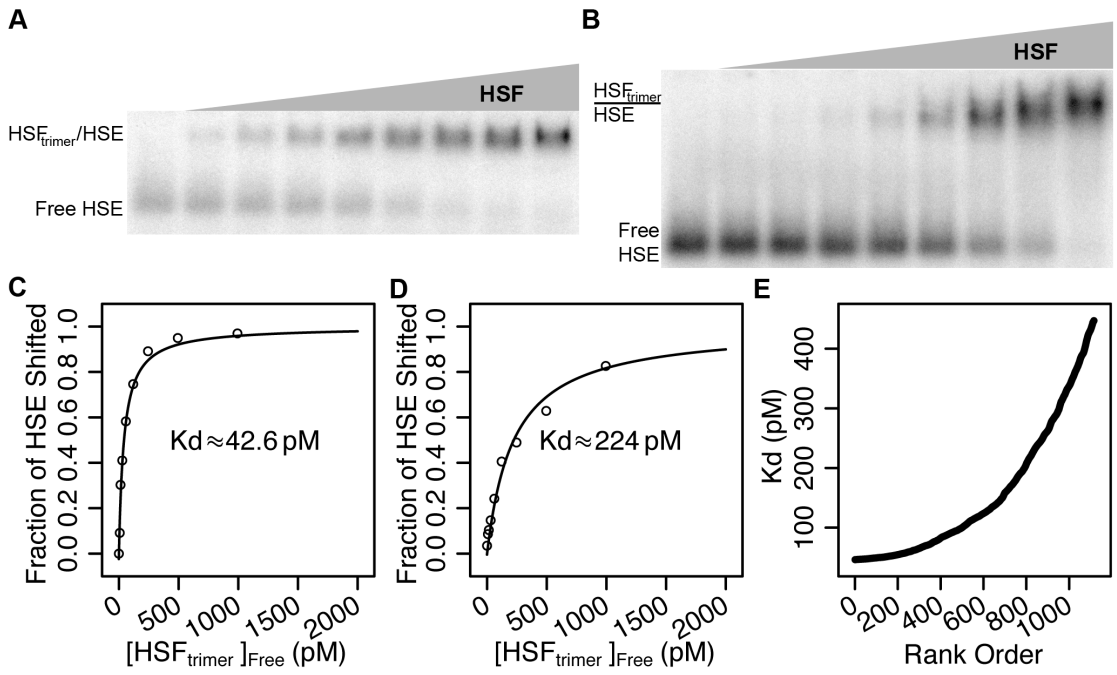

Figure 2.2: Recombinant HSF binds HSEs with picomolar affinity in vitro. A and B) The mobility of the constant 200 attomole HSE probe shifts into a trimeric-HSF:HSE complex as increasing HSF is added. There is no HSF in the left-most lane, the right-most lane contains 3 nM HSF (1 nM trimeric HSF), and the intervening lanes contain two-fold serial dilutions of HSF. C) A hyperbolic curve based on the Kd equation (see Methods) was modeled using the band shift data, indicating a Kd of 42.6 pM (95% confidence interval of 36.8-49.4 pM). D) A hyperbolic curve based on the Kd equation (see Methods) was modeled using the band shift data, indicating a Kd of 224 pM (95% confidence interval of 181-276 pM). E) The intensity of each isolated HSE in the *Drosophila* genome is transformed to an absolute Kd using the absolute Kds calculated from band shift data in panels A and B. The Kd values range from 40-400 pM.

(Figure 2.2E and Methods). We confirmed the transformation of the relative Kd values to absolute Kds by performing band shifts with genomic HSEs of different predicted Kd values (Figure A.2). The experimental verifications of the measurements are within the estimated error of the EMSA confidence interval and the variability between PB-seq replicates (Figure A.3).

Taken together, these measurements allow us to characterize the binding energy landscape for HSF across the entire *Drosophila* genome, in the absence of chromatin. Our estimated Kd values for isolated HSEs in the *Drosophila* genome

ranged from 40-400 pM (Figure 2.2E). These in vitro binding results demonstrate the feasibility and efficiency of combining high-throughput detection methods with classic EMSA and competition experiments to quantify the binding energy for the comprehensive set of potential genomic binding sites for a sequence-specific TF.

2.2.2 Chromatin features and PB-seq data predict HSF binding intensity in vivo

Our data reveals substantial differences between in vivo and in vitro binding intensities (Figure 2.3A), underscoring the role of chromatin in determining in vivo binding site selection and affinity. We found DNase I hypersensitivity was the most important predictor of HSF binding; therefore, we scaled the in vivo and the in vitro read counts so that they were approximately equal at in vivo sites with high DNA accessibility (Methods, Figure A.4). After this normalization, we partitioned the binding sites that were detectable in vitro into classes: "unaffected" sites, bound at comparable affinities in vivo and in vitro (55 red points in Figure 2.3A; 2% of all sites); "suppressed" sites, with reduced, but detectable, in vivo intensity (365 green points; 13%); and "abolished" sites, below the in vivo threshold for detection (2223 blue points; 76%). In addition, sites not detectable in vivo or in vitro were labeled "background" (249 gray points; 9%), and sites with stronger relative in vivo intensity compared to in vitro were labeled "enhanced" (4 black points; 0.1%).

PB-seq data reveals potential HSF binding sites, providing the opportunity to model the effect that non-stressed chromatin landscape has upon induced

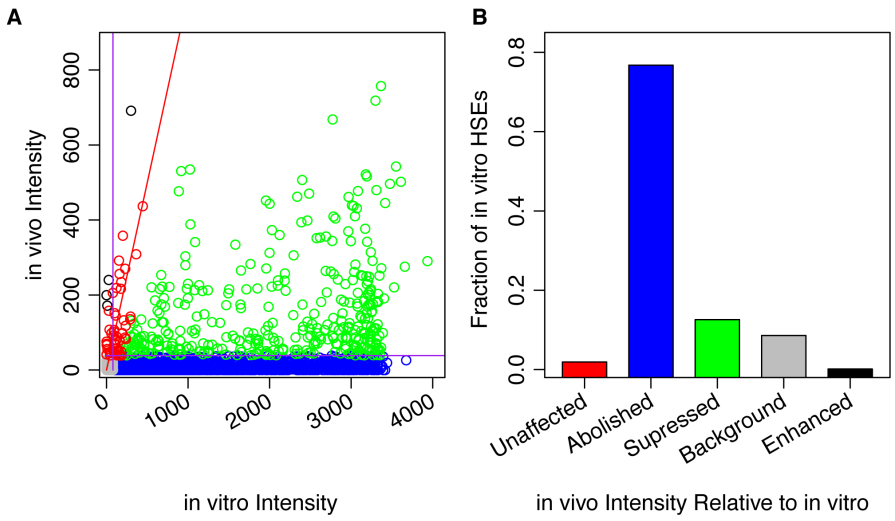

Figure 2.3: In vitro and in vivo binding of HSF to genomic HSEs do not correlate. A) A scatter plot comparing the observed in vivo HSF binding intensity and in vitro binding intensity for each isolated HSE indicates that the vast majority of in vivo binding is suppressed (green) or abolished (blue), if we assume that the top seven most DNase I hypersensitive isolated HSE clusters provide the best estimates for sites that are minimally influenced by chromatin. After scaling, red points have similar in vivo and in vitro intensity, black points may be enhanced in vivo, while green and blue points are suppressed and abolished, respectively. B) The points from panel A were categorized, and the resulting bar chart shows the relative frequencies of each category.

HSF binding intensity. There is a wealth of chromatin data available for S2 cells during unstressed conditions [77][47], and heat-shock induced binding sites of HSF in S2 cells are also known [53]. We used DNase I hypersensitivity data [77], MNase data [47] and ChIP-chip data for 9 factors and 21 histone modifications for unstressed *Drosophila* S2 cells (Table A.1) [77] to predict the intensity of inducibly bound in vivo HSF-bound sites (Figure 2.4A, Figure A.5 and Figure A.6). For our statistical model, we selected a *rules ensemble* [43], a linear regression model in which some terms are combinations of covariates known as "rules". This approach allowed us to capture fairly complex interactions between covariates. For example, a rule might apply when H3K27ac and DNase I hypersensitivity both exceeded designated thresholds (value ranges can also be expressed). Each rule's coefficient is added to the predicted value if, and only if,

the rule applies. When there is only one covariate, the model reduces to a linear regression.

The Pearson's correlation coefficient between HSF ChIP-seq data for the model incorporating all these data sets was r = 0.62 (Figure A.6 and Figure A.7). As the large number of covariates brings with it some danger of overfitting, we tested combinations of the four classes of covariates: DNase I hypersensitivity, MNase, histone modifications/variants, and non-histone factors (Figure 2.4B, Figure A.6, Figure A.7). Of notice, the correlation of the linear regression model that incorporates DNase I data was r = 0.64 on the test data (Figure 2.4B and Figure A.7B). Our study is consistent with a previous study that obtained r = 0.65 for actual and inferred TF binding intensities using a DNase I dependent model [73].

Other covariate classes produce similar, but lower, correlations. The rules model using histone modifications and histone variants yielded r = 0.57 (Figure 2.4B and Figure A.7), while a rules model incorporating non-histone bound chromatin factors yielded r = 0.58 (Figure 2.4B and Figure A.7). Combining covariate classes further improves the correlation to as much as r = 0.70 (Figure A.6 and Figure A.7). We also examined the Receiver Operator Curves (ROC) for the different covariate combinations (Figure A.8) and found concordant results. If we assume that the PB-seq, genomic ChIP, DNase I-seq, and MNase-seq experiments are maximally resolved and sensitive, with no experimental noise, an approximate upper bound is given by r = 0.90, as observed for two HSF-ChIP-seq replicates [53]. Notably, the higher resolution of the DNase I-seq data, compared to the ChIP-chip data, may be why DNase I-seq alone is strongly predictive in the linear regression model and most influential in the rules ensemble

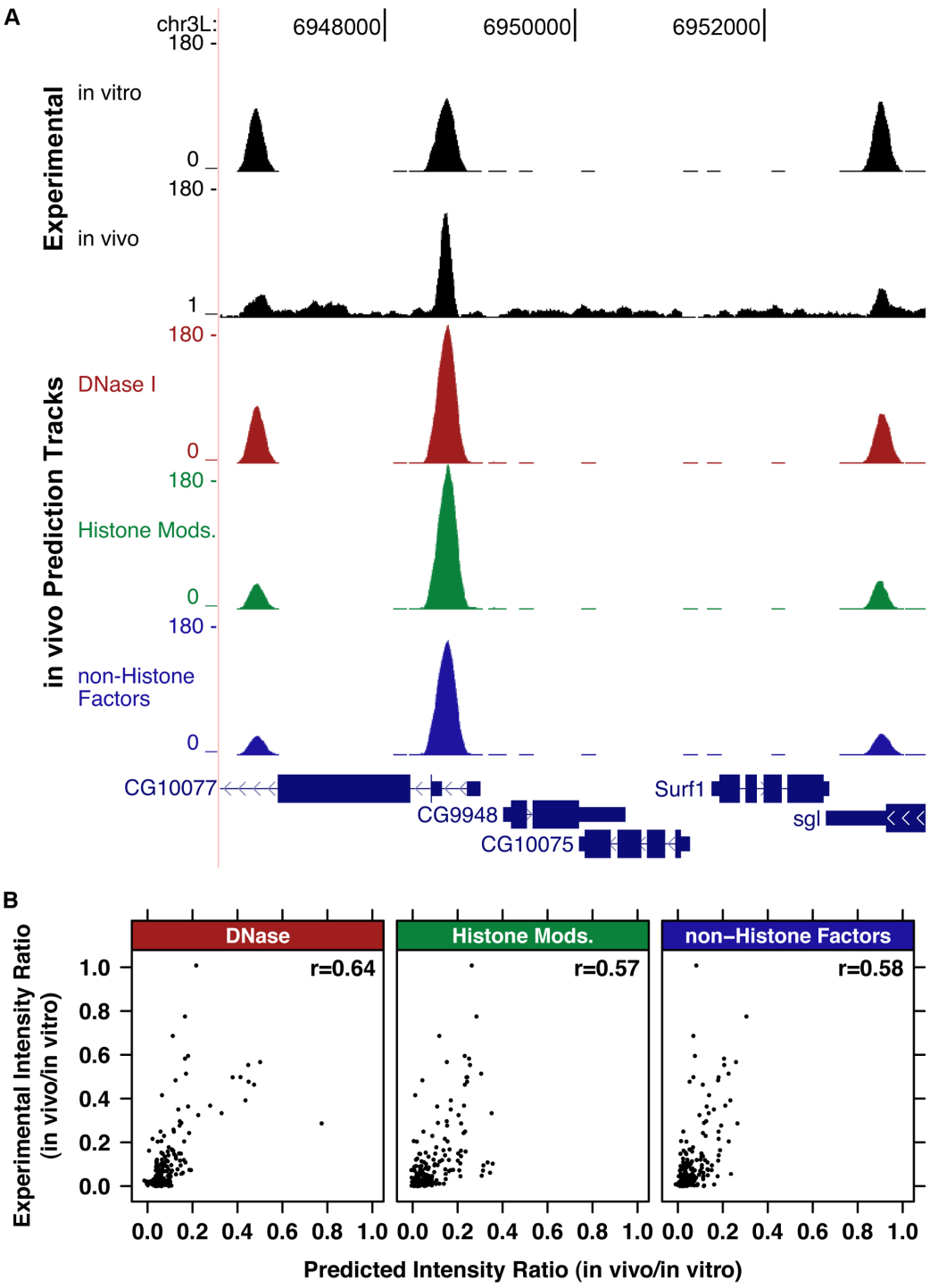

Figure 2.4: Genomic chromatin and PB-seq data accurately predict in vivo HSF binding intensity. A) The intensity of in vivo ChIP-seq peaks is not recapitulated by in vitro PB-seq data; however, genomic DNase I hypersensitivity data and histone modification ChIP-chip data can be used to accurately predict HSF binding intensity. B) The experimentally determined ratio between in vivo ChIP-seq HSF intensity and in vitro PB-seq intensity is plotted against the predicted in vivo/actual PB-seq ratio. The Pearson correlation for each model is shown.

models. Notably, we used the chromatin landscape prior to induced TF binding to predict binding intensity, whereas previous models have used the chromatin landscape present when the TF is bound in order to infer binding intensity [73] or infer binary binding events [106][18] (see Discussion).

Our data and modeling indicated that the presence of active chromatin features, such as histone acetylation and DNase I hypersensitivity, had a significant influence on the predictive power of the model, while repressive features had minimal influence (Figure A.9). DNase I hypersensitivity was a strongly predictive covariate in the model when used in a simple linear regression model (Figure 2.4), or in combination with histone modification and non-histone factor covariates in the rules (Figure A.9 E- A.9G, A.9J, A.9K and A.9M). Tetra acetylation of H4 and H3K9ac were the most informative histone marks in the model that used histone variants and histone modifications as covariates (Figure 2.5A). GAGA associated factor (GAF), which has a proposed role in permitting HSF binding [82], was the most influential factor in the HSF binding prediction model that considered all chromatin-binding factors (Figure 2.5B).

2.2.3 Defining genome-wide DNA accessibility by chromatin composition

The analysis above indicates that DNA accessibility, as measured by DNase I hypersensitivity, is a primary determinant of binding intensity. Previous studies have similarly shown that TF binding sites correlate strongly with DNase I hypersensitive sites [73][106][18][68]. For instance, histone acetylation causes local chromatin decondensation by reducing the ionic interactions between lysine

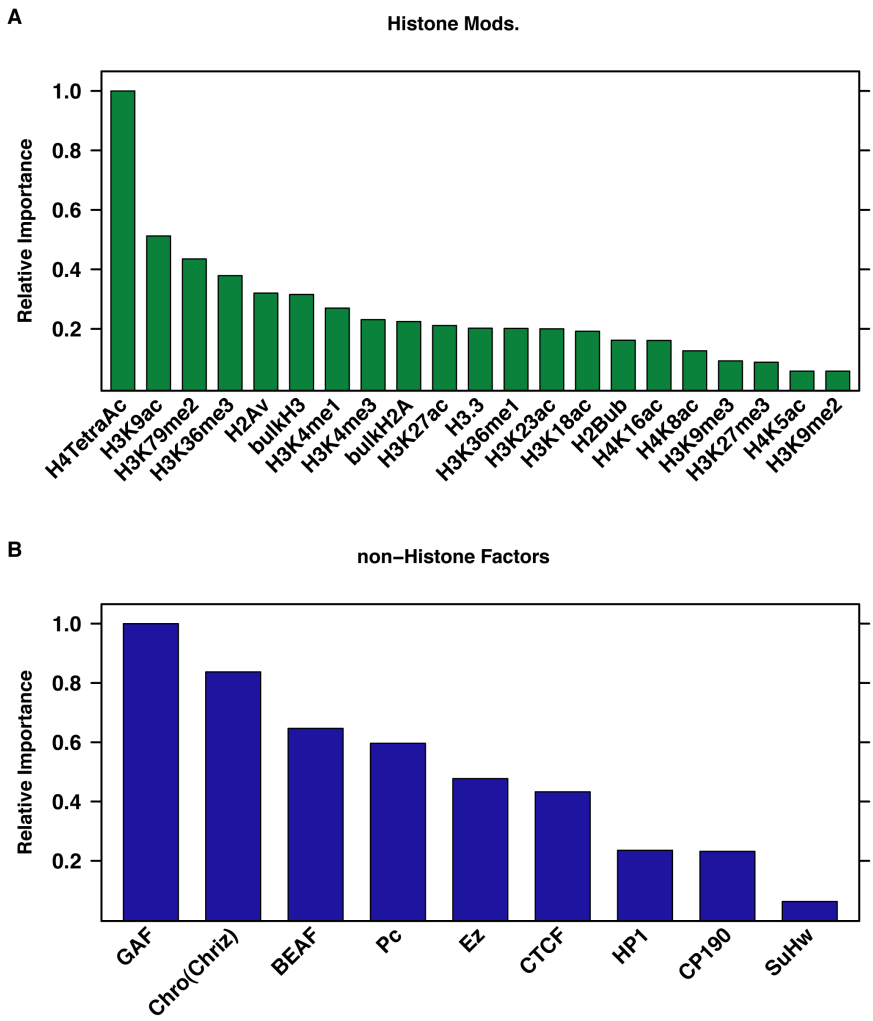

Figure 2.5: Histone acetylation and GAF occupancy are important covariates in predicting HSF binding intensity. Plotted are the relative values of the sums of the coefficients associated with all rules that reference each covariate in the rules ensemble [43]. Results are shown for (A) the histone variant and modification model and (B) the non-Histone factor model.

residues and DNA and promotes accessibility, but the extent to which combinations of histone marks and TFs act together to dictate chromatin accessibility is not known. Therefore, it is of interest to see whether DNA accessibility can be predicted from specific features of the chromatin landscape, such as histone modifications and non-histone chromatin bound factors. In addition, accurate predictions of DNA accessibility would be of practical use, because direct measurements are often not available.

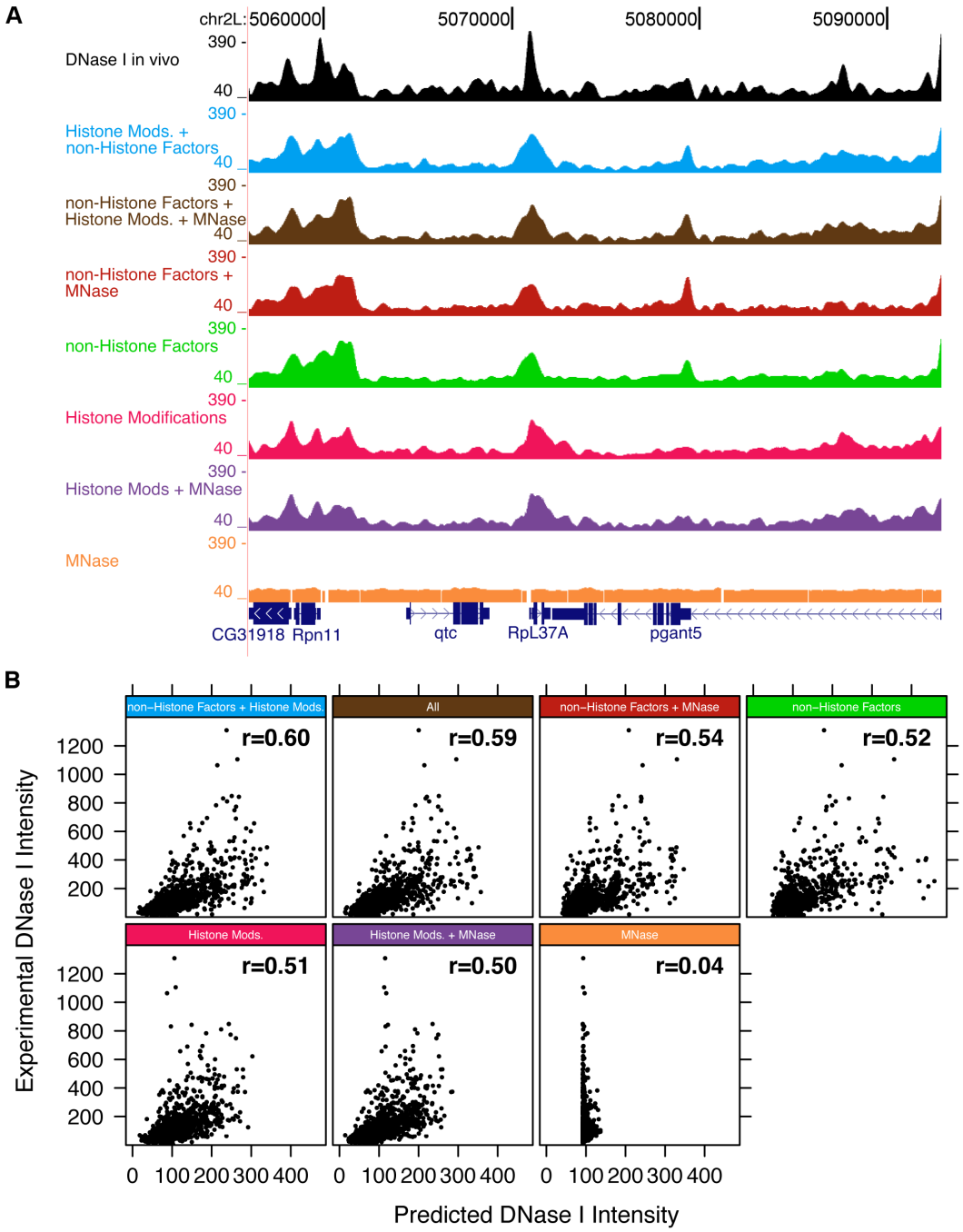

Figure 2.6: DNase I hypersensitivity can be inferred using histone marks and MNase data. A) The intensity of DNase I hypersensitivity landscape is inferred by models (colors) that use histone modification profiles, non-histone factor profiles, DNase I data and MNase-seq data. B) The experimentally determined DNase I hypersensitivity data is plotted against inferred intensity for the various models. The Pearson correlation for each model is shown.

To address this question, we applied our rules ensemble framework to predict DNase I hypersensitivity (the best available proxy for DNA accessibility) from ChIP-chip data for histone features, non-histone chromatin bound factors, MNase data and combinations of these covariate pools (Figure 2.6). Tetra-acetylation of H4 and H3K9 acetylation were most influential in the model that uses histone modifications, bulk histone and histone variant intensities (Figure A.10E); the correlation coefficient for this model using the test data is 0.51 (Figure A.11B). The model that uses non-histone factor ChIP-chip data obtains a correlation of 0.52 (Figure A.11B), which is consistent with TFs having characteristic DNase I hypersensitivity footprints [106][18]. The model that combines both histone data and non-histone data into a rules model performs the best on the test set, with a correlation of 0.60 (Figure A.11B). Repressive histone marks appear to contribute little to generating the DNase I hypersensitivity pattern (Figure A.10) and the lack of active chromatin marks appears to be sufficient to package DNA into inaccessible units. These models reinforces the notion that the biochemical composition of chromatin permits DNase I hypersensitivity and quantifies the contributions individual modifications, and combinations thereof, make to DNase I hypersensitivity (Figure A.11). As more and higher-resolution genome-wide data becomes available, these models will be refined.

2.2.4 Dissection of the Heat Shock Element

PB-seq provides the opportunity to model the sequence-dependent binding preferences of a purified TF genome-wide and independent of chromatin or other factors. In the case of HSF, the consensus binding site is well characterized and consists of three pentamers, "AGAAN NTTCT AGAAN", (here denoted

pA, pB, and pC), each bound by a monomer of the HSF homotrimer. Note that the consensus sequences for pA and pC are identical, while the one for pB is their reverse complement. Of course, the consensus HSE is a crude summary that ignores subtleties in the base preferences at each position. A position-specific scoring matrix (PSSM) provides a somewhat improved description but still ignores dependencies between positions within the binding site. We sought to use genome-wide binding sites from PB-seq to produce an improved model for the sequence preferences at HSEs.

We began by computing the mutual information for all pairs of HSE positions based on the identified in vitro binding sites. We found negligible evidence of correlated base preferences between positions, but we did observe that some pentamers within PB-seq peaks adhered closely to the consensus motif while others did not. This led us to formulate a probabilistic model that allows each pentamer in an HSE to closely match the consensus ("strict") or diverge from it more substantially ("relaxed"), and considers all possible combinations of pentamer composition (Figure A.12). More specifically, we described each of the three pentamers using a two-component mixture model, with a latent variable indicating "strict" or "relaxed" binding preferences, and estimated the joint distribution of these three latent variables from the data.

The model parameters — the position-specific nucleotide probabilities and prior distribution for the combinations of strict/relaxed pentamers – were estimated from the data by maximum likelihood using an expectation maximization algorithm (see Methods). In fitting the model, we considered only the 1309 isolated HSEs, sequence elements that were at least 200 base pairs away from any other degenerate HSE motif, to avoid complications arising from overlap-

ping HSEs. The model fit the data substantially better than did a simple PSSM (lnL = 15442 vs. lnL = 15673 for the PSSM; Akaike information criterion [AIC] = 15636 vs. AIC = 15763 for the PSSM), suggesting that it effectively captures important dependencies between positions.

Based on the estimated model parameters, we computed a posterior probability distribution over all combinations of pentamer stringency and order for each HSE (Methods; Figure 2.7B). These values were averaged across HSEs to obtain expected genome-wide fractions of HSEs having each of the strict/relaxed pentamer combinations. We found that binding sites with strict pB and pC, and relaxed pA, were most frequent (an expected 38% of sites), indicating that this configuration is preferred (Figure 2.7B). The next most frequent configurations were a relaxed pB flanked by a strict pA and pC (33%), and a strict pA and pB combined with a weak pC (29%). Interestingly, combinations of three strict pentamers occur at negligible frequency. Indeed, only 5 out of 1309 isolated genomic HSEs matched the consensus sequence exactly, while 148 differed from it by a single mismatch. Configurations with at most one strict pentamer were also rare. Together, these results indicate that the biophysical interactions of the pentamers within the binding sites are critically dependent upon their composition and position relative to the other pentamers in an HSE.

While the three estimated strict pentamer matrices were similar (Figure 2.7A top), the relaxed matrices showed substantial differences with respect to each other (Figure 2.7A bottom). For example, the relaxed pA matrix indicates that 70-80% of HSEs containing a weak pA have the consensus base at positions two, three and four. In contrast, position 12 in pC (the analog of position 2 in pA) almost invariably contains a G in all HSEs, while positions 7 and 8 in

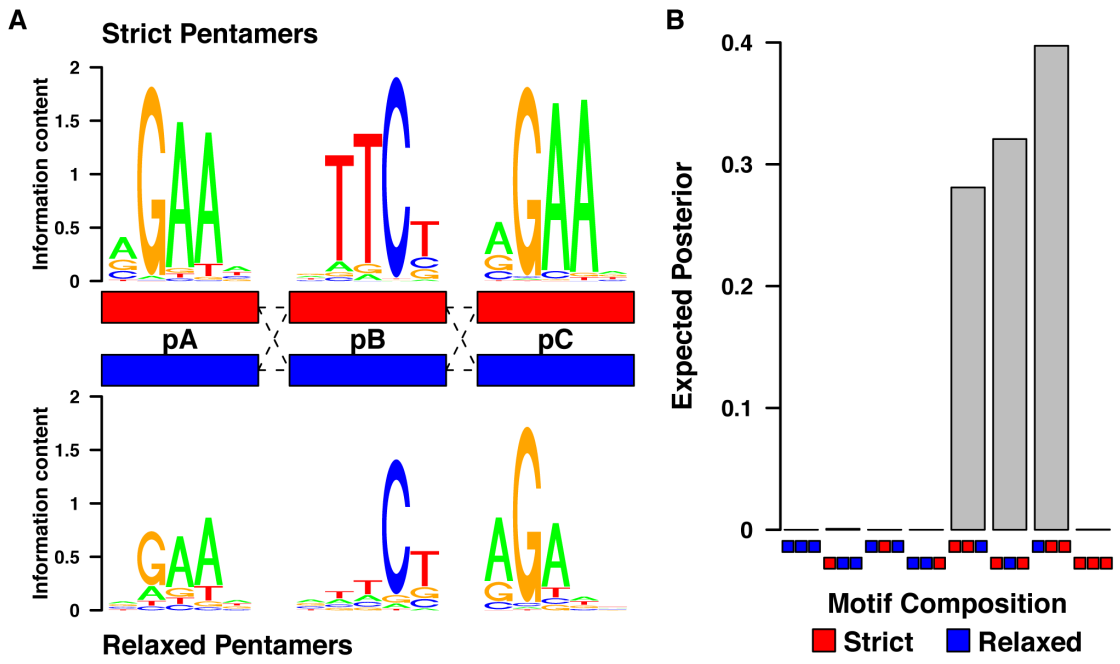

Figure 2.7: Pentamers within the HSEs are dependent upon their consensus match and also their position relative to the other pentamers. A) The mixture model defines each pentamer within the HSE as strict or relaxed depending upon how well it conforms to the canonical HSE. Note that the position of relaxed pentamers strongly influences their composition. B) A probabilistic sequence model reveals that the presence of two strict (red) and one relaxed (blue) pentamer provides the best explanation of the data.

pB (analogous to positions 3 and 4 in pA) have only modest base preferences in HSEs containing a weak pB. This analysis indicates that each monomeric HSF/pentamer interaction has distinct biophysical properties within the context of the broader HSF/HSE interaction. We also devised a simplified model, with a single strict matrix shared by all three pentamers, and a single relaxed matrix obtained by applying a "dampening" factor to the strict matrix (Figure A.13, Methods). This model further supports the strict/relaxed pentamer split (lnL = 15908 vs. lnL = 16048 for a single-monomer PSSM; and AIC = 15952 vs. AIC = 16078), although both the full model and the full PSSM fit the data better (lower AIC). Moreover, not only was the simplified model still able to reproduce the posterior distributions over pentamer configurations of the full

model, but it was also able to replicate synthetic patterns from simulated data (Figure A.14). Finally, the preference for single pentamer degeneracy was also observed independently by comparing the pentamer-specific KL-divergence in PSSMs obtained from subsamples of HSF bound peaks (Figure A.15; Methods).

2.3 Discussion

The PB-seq technique combined with EMSA and competition assays provides a straightforward, yet versatile and powerful framework for characterizing all potential binding sites in a genome, regardless of tissue specificity, developmental stage, or environmental conditions. Comparing in vitro and in vivo binding profiles, in the context of pre-induction genomic chromatin landscape, revealed DNase I hypersensitivity, H4 tetra-acetylation, and GAF as critical features that modulate cognate element binding intensity in vivo. Furthermore, DNase I sensitivity was found to be strongly influenced by high GAF occupancy and histone acetylation, while repressive factors were minimally influential in the statistical models. Finally, the full set of potential genomic binding sites provided a rich data set that was used to build more detailed sequence models, which tease apart substructure and features that are lost with traditional PSSM models.

One initially surprising observation from our study was that 40% of the in vivo HSF peaks were not found in vitro. We believe that the limited dynamic range for quantifying in vitro binding affinity may be responsible for the lack of detectable in vitro peaks. Although we quantify in vitro binding over an order of magnitude (40-400 pM), the experimental concentrations of HSF and genomic DNA and our depth of sequencing do not permit the detection of lower affin-

ity HSF binding sites. For instance, only eleven sequence tags would be predicted to underlie a hypothetical 5 nM HSF binding site, and these would not be distinguishable from background. Upon further examination, we find that the composite HSE representing those in vivo binding sites that were not found in vitro is more degenerate than those found using both assays (Figure A.16A). Moreover, the in vivo sites that were not found using PB-seq were also more accessible in vivo (Figure A.16B), in support of our hypothesis. Performing PB-seq at a range of protein and DNA concentrations, or increasing sequence coverage would expand the dynamic range of quantification by PB-seq.

Other possible explanations for this observation include cooperative interactions with pre-bound chromatin factors, long-range DNA interactions, post-translational modifications of HSF, higher-order chromatin structure, or bridging protein interactions. The influence of DNA modifications and immediate flanking sequence do not contribute to this disparity, since we use large fragments of purified genomic DNA. Bridging protein interactions [49], which do not involve HSF directly binding to DNA, appear not to be responsible for our results because 95% of in vivo peaks encompass at least one HSE near the peak center [53]. However, if other proteins were cooperating with HSF in vivo to enhance HSF binding intensity at low affinity binding sites, then some of these peaks may not be observed in vitro. Since our PB-seq experiment used recombinant HSF in the binding experiments, we would also not capture differences in binding site affinities that are due to post-translational modifications of HSF [22]. To overcome these potential limitations, PB-seq could be adapted to include known bridging/cooperative factors and proteins could be purified from in vivo sources to capture indirect or modification-dependent interactions.

The notion that motif accessibility is driving inducible TF binding in vivo is supported by independent studies of distinct TFs: STAT1, HSF, glucocoticoid receptor (GR), and GATA1 [53][68][117][145]. These studies show that the chromatin landscape prior to TF binding influences inducible TF binding. In the first study, it was found that a large fraction of STAT1 induced binding sites contained H3K4me1/me3 marks prior to interferon-gamma (IFN-) induced STAT1 binding [117]. Our group previously found that inducible HSF binding sites are marked by active chromatin compared to sites that remain HSF-free [53]. A more recent study has shown that inducibly bound GR sites are marked by DNase I hypersensitive chromatin prior to GR binding [68]. Likewise, the permissive chromatin state at GATA1 binding sites is established even in GATA1 knock out cells [145]. While these correlations are instructive, no previous attempt has been made to model inducible TF binding using biological measurements of chromatin landscape present prior to TF binding. Recent models have successfully inferred TF binding profiles using DNA sequence and chromatin landscape data, generated at the same time the TF is bound [73][99][106][18]. However, these models do not distinguish between the influence TFs have upon local chromatin and the chromatin features that permit TF binding. In contrast, we modeled the changes between HSF in vitro binding (PB-seq) and in vivo binding (ChIP-seq) landscapes as a function of the non-heat shock chromatin state. This produced a quantitative model describing the important features that modulate the in vivo HSF binding intensity. Moreover, the use of our rules ensemble model enabled the capture of potential interactions between these chromatin features.

Our study reveals that DNase I hypersensitivity and acetylation of H4 and H3K9 are strong predictors of inducible HSF binding intensities, however the

molecular events and factors that precede TF occupancy to maintain accessible chromatin remain poorly characterized. For instance, the degree to which pioneering factors or flanking DNA sequence, individually or in combination, maintain or restrict accessibility remains unclear. A recent study highlights the biological consequences of maintaining the inaccessibility of TF binding sites, in order to repress expression of tissue-specific transcription factors in the wrong tissues. The authors found that ectopic expression of CHE-1, a zinc-finger TF that directs ASE neuron differentiation, in non-native *C. elegans* tissue is not sufficient to induce neuron formation [136]. However, combining ectopic CHE-1 expression with knockdown of *lin-53* did modify the expression patterns of CHE-1 target genes in non-native tissue, effectively converting germ line cells to neuronal cells [136]. Lin-53 has been implicated in recruitment of deacetylases, and deacetylase inhibitor treatment mimics *lin-53* depletion, suggesting that Lin-53 is actively maintaining CHE-1 target sites inaccessible in germ cells.

Alternatively, functional TF binding sites could be actively maintained in the accessible state. HSF binding within ecdysone genes has a functional role in shutting down their transcription [48], and activating ecdysone-inducible genes containing inaccessible HSEs causes chromatin changes that are sufficient to allow HSF binding [53]. In this special case of HSF-bound ecdysone genes, active transcription and the corresponding histone marks are mediating access to HSEs, in order for HSF to bind and repress transcription upon heat shock. A more recent study has shown that activator protein 1 (AP1) actively maintains chromatin in the accessible state, so that GR can bind to cognate elements [16].

Although TF accessibility to critical genomic sites appears to be actively maintained, many binding sites may be a non-functional result of fortuitous

TFBS recognition. It has long been hypothesized that the binding affinities for TF/DNA interactions are sufficiently strong to allow promiscuous binding at the cellular concentrations of TFs and DNA [140][87]. There are roughly 32,000 HSF molecules per tetraploid S2 cell [44] and the dissociation constants for trimeric-HSF/HSE interactions are in the picomolar range (Figure 2.2E); therefore much of the in vivo HSF binding may be non-functional promiscuous binding. Additional investigation will further illuminate the role of chromatin context in TF binding and the mechanisms by which programmed developmental or environmental chromatin changes permit or deny TF binding.

Elucidating the rules that govern accessibility is essential for predicting in vivo occupancy of TFs. Diverse transcription factors [86], from a broad spectrum of organisms [68], bind their sequences based on site accessibility. We found that chromatin accessibility as measured by DNase I hypersensitivity could be inferred using ChIP-chip data for various histone modifications and transcription factors. Although our model can infer accessibility based on chromatin composition, the mechanism by which accessibility originates is not addressed. Previous studies have shown that activators, such as HSF, glucocorticoid receptor, and androgen receptor bind to their cognate sites and direct a concomitant increase in local acetylation, DNase I hypersensitivity, and nucleosome depletion [53][68][141][59]. Androgen receptor also acts to position flanking nucleosomes marked by H3K4me2 [59]. These post-TF binding chromatin changes that occur are the result of acetyltransferase and nucleosome remodeler recruitment, both of which functionally interact with activators. For instance, both GR and GATA1 interact with the nucleosome remodeling complex Swi/Snf [65][67]. Concomitant increases in locus accessibility likely allow large molecular complexes such as RNA Pol II and coactivators to access the

region that in turn can reinforce and maintain active and accessible chromatin.

Thorough biophysical characterization of TF binding site properties is critical for accurate predictions of TF binding sites, underscoring the need for more complete models of TF binding. While the commonly used PSSM model makes the assumption of base independence, recent work has revealed that richer models providing for interactions between positions are necessary [124][60]. Our model captures critical features of the HSF/HSE interaction that are lost with simpler computational models, namely the interdependencies between the sub-binding sites of each HSF monomer. Consistent with our model, a series of in vitro experiments with *S. cerevisiae*, *D. melanogaster*, *A. thaliana*, *H. sapien* and *D. rerio* HSFs indicate that HSF from each of these species can bind to discontinuous HSEs containing canonical pentamers that contain intervening five base pair gaps [36][120]; interestingly, however, *C. elegans* HSF strictly binds to continuous HSEs that do not contain gaps [36]. The complex interactions between positions within a binding site are a critical aspect of inferring whether a polymorphism or mutation affects TF binding. These features should prove useful in providing degenerate HSE sequences for optimal co-crystallization of trimeric HSF and DNA and inferring changes in DNA sequence that affect HSF binding within and between species.

In conclusion, the data and models presented here reinforce both the importance of chromatin landscape in modulating in vivo TF binding intensity and how genome wide, chromatin free, binding assays contribute to the understanding of TF sequence binding specificity.

2.4 Methods

2.4.1 Cloning and purification of recombinant HSF

Drosophila HSF was N-terminally tagged with glutathione s-transferase and a tobacco etch virus (TEV) protease cleavage site. The C-terminus of the recombinant HSF was fused to the 3xFLAG epitope. Recombinant HSF was purified from *E. coli* with glutathione resin as previously described H:2010eh, with the following modifications: HSF-3xFLAG elution was achieved by addition of 6xHistidine tagged TEV protease and TEV protease was cleared from the HSF preparation using a Nickel-NTA column. Densitometry was used to show that the HSF protein preparation was 40% full length HSF-3xFLAG, and known amounts of bovine serum albumin (BSA) were used to quantify the HSF (Figure A.1).

2.4.2 Band shift assay

Serial two-fold dilutions of recombinant HSF, from 3 nM (1.5 nM for the 221 pM HSE) to 23.3 pM, was incubated with 200 attomoles of radiolabeled dsDNA containing modestly degenerate HSEs (chrX:3380775-3380824 (224 pM), chr2L:5009892-500994 (42.7 pM), chr2R:3529792-3529841 (308 pM), chr3L 13470978-13471009 (221 pM), and chr3L:4073542-4073591 (97.5 pM)) and allowed to come to equilibrium for 30 minutes in a total of 10 μl of 1xHSF binding buffer (20 mM HEPES pH 7.9, 10% glycerol, 1 mM EDTA, 4 mM DTT, 3 mM MgCl2, 100 mM NaCl, 0.1% NP-40, and 300 μg/ml BSA) at room temperature. Binding reactions were loaded in a 3% agarose TBE (10 mM Tris-HCl pH 8.0, 25

mM boric acid, and 1 mM EDTA) gel and electrophoresed at 50 Volts for 2 hours. The HSF-bound probe and free probe were quantified by densitometry and the dissociation constant, $K_d = ([A][B])/[AB]$, was estimated using a non-linear least squares method on the function $[AB]/[A]_{total} = [B]/([B] + K_d)$ where $[AB]/[A]_{total}$ is the measured shifted fraction and [B] is the free HSF trimer concentration.

2.4.3 PB-seq: Genomic in vitro binding experiment

We incubated 600 pM HSF and 2500 ng genomic DNA (sonicated to 100-600 bp fragment size as previously described [53]) in 1500 µl final volume of 1xHSF binding buffer and let it come to equilibrium for an hour at room temperature. We added 20 l ANTI-FLAG M2 affinity gel for 10 minutes and washed 8 times with 1xHSF binding buffer to remove unbound DNA, 3xFLAG peptide was added to a final concentration of 200 ng/µl to specifically elute HSF and HSF-bound DNA. The mock IP was done in the absence of recombinant HSF. We attribute the in vitro binding assay's low background to the design of the experiment. Since recombinant C-terminally 3xFLAG tagged HSF was used, the HSF-associated DNA could be specifically eluted by the addition of excess 3xFLAG peptide. In contrast, standard ChIP protocols rely on non-specific elution of all protein and DNA that binds the resin.

2.4.4 Illumina library preparation

The sample preparation was as previously described [53], except that 15 rounds of amplification were performed in this case.

2.4.5 PB-seq HSF peaks and HSE sites

The PB-seq reads were aligned to the *Drosophila* Genome (BDGP R5/dm3) using BWA (v 0.5.8c) [85]. We obtained 5,052,425 uniquely aligned reads for replicate one, 4,694,846 for replicate two and 5,410,049 for the mock. Files that contain raw sequence data and uniquely aligned reads were deposited into NCBI's Gene Expression Omnibus (GEO) [10], accession number GSE32570.

We called peaks using MACS (v 1.3.7.1) [149], both for each individual replicate and for the merged set, using a tag size of 55 bp, a starting bandwidth of 100 bp and an appropriate genome size. After experimenting with several p-value thresholds, we selected a value of p = 0.01, which achieved a good tradeoff between maximizing the number of called peaks and ensuring consistency between replicates. Our results were largely unaffected by the "mfold" parameter (the threshold for fold enrichment relative to background for inclusion in the peak model), so we left this parameter at its default value.

To improve our sensitivity in binding site detection, we made use of an ensemble of position weight matrices (PSSMs), rather than a single matrix. We sampled 10,000 sets of 100 peaks and used the program MEME [7] for motif discovery in each set. As input, MEME was given the 100 bp sequence centered at each peak summit. We used a fixed motif width of 14 bp, a second order background Markov model estimated from the entire peak set, and the "zoops" model (zero or one site per sequence) with the restriction that at least 75% of the sequences must contain a site. The resulting PSSMs were compared by KL-divergence against the canonical monomer PSSM (four base pair unit with consensus AGAA) estimated from the previously published in vivo high-confidence HSF binding sites detected by ChIP-seq [53]. In each PSSM, one of

the three monomers had on average about twice the KL-divergence as the other two. Figure A.15 shows a scatter plot of the KL-divergence of the PSSMs in the ensemble.

Each peak was scanned for matches to all PSSMs in the ensemble, allowing for overlapping sites. The score at each position was taken to be the maximum score across the ensemble. Peaks were split into three groups by GC% quantile, and for each group a 10 kbp sequence was simulated from a second order Markov model, which was then used to estimate the FDR associated with the score.

In our context, an appropriate FDR threshold should strike a balance between recapitulation of in vivo results and limiting the number of spurious binding sites. In vivo results are defined by high-confidence peaks, which are ChIP-seq peaks that were called by two peak calling programs and have a corresponding binding site sequence underlying the peak [53]. Whereas, spurious sites are accounted for by limiting the average number of HSE clusters per peak (set of potentially overlapping HSE no more than 10 bp apart from each other). Due to the repetitive nature of the HSE, a cluster is a better representative than a single site of a functional binding locus. We chose a 20% FDR threshold, which maximizes the fraction of peaks having a single HSE cluster while ensuring that a large fraction (97%) of the high-confidence in vivo peaks contain HSEs. This threshold resulted in 3735 clusters (71% with a single HSE, 20% with two HSEs overlapping by 10 bp, ~ 5% with two HSEs overlapping by 5 bp; see Figure A.17).

The final set of HSE clusters was obtained by combining data from the two experimental replicates. First, a set of genomic regions was identified by inter-

secting the peaks from the two experimental replicates, and retaining only those peaks for which the two replicates were in close agreement (>80% of reads fall in the overlapping region). We then identified the 2896 HSE clusters that fell in these regions (~ 77% of all clusters).

2.4.6 HSE cluster intensity

The problem of measuring the intensity of each peak is complicated by the fact that some peaks contain multiple, closely spaced clusters, whose contributions are difficult to disentangle. Furthermore, peaks often include trailing edges that are dominated by the background signal. To address these concerns we experimented with various measures of intensity based on the output produced by MACS (wig files giving shifted read counts in 10 bp windows) as well as the reported "bandwidth" B. We considered three measures, applied to a window of radius B centered at each cluster: maximum read count, read count sum, and an "integrated" read count based on a biweight kernel (which produces a curve at each peak that is similar to the one implied by the peak model used by MACS). We selected the biweight kernel measure, which does the best job of handling closely spaced clusters (see Figure A.18).

2.4.7 Computing K_d values for all genomic HSE sites

We assume that each HSE site i is at approximately the same initial concentration in the experiment ($[HSE_i]^{initial} = C$). Furthermore, all sites compete to bind a shared amount of free HSF, with the remaining unbound concentration de-

noted by [HSF]. At the end of the experiment, a fraction of site i is bound, with concentration $[HSE_i : HSF]$, and the remainder is unbound, with concentration $[HSE_i]$. The dissociation constant for a particular HSE site is therefore given by:

$$K_d^i = \frac{[HSF][HSE_i]}{[HSE_i : HSF]}$$

The bound HSE concentration is measured by the PB-seq experiment in terms of the number of reads at element i (R_i). This leaves two unknown quantities, $[HSF]$ and $[HSE_i]$, in units of read counts. The first of these unknowns, $[HSF]$, can be eliminated by considering instead the relative K_d with respect to a known reference value (for an HSE present in the experiment). To solve for $[HSE_i]$, we express this quantity as the difference between the initial concentration C and the measured bound concentration:

$$[HSE_i] = [HSE_i]^{initial} - [HSE_i : HSF] \propto C - R_i$$

By substituting the expression for K_d (above) and dividing by the K_d value for the reference HSE, K_d^{ref} we obtain an expression with a single unknown, C:

$$\frac{K_d^i}{K_d^{ref}} = \frac{\frac{[HSF][HSE_i]}{[HSE_i:HSF]}}{\frac{[HSF][HSE_{ref}]}{[HSE_{ref}:HSF]}} = \frac{\frac{[HSE_i]}{[HSE_i:HSF]}}{\frac{[HSE_{ref}]}{[HSE_{ref}:HSF]}} = \frac{(C - R_i)R_{ref}}{(C - R_{ref})R_i}$$

With the use of a reference dissociation value for a second HSE, we can solve for C and obtain estimates of the dissociation constants for all other HSE sites for which read counts are available. Replacing K_d and R_i by the corresponding values for the second reference HSE and solving for C:

$$K_d^{ref2} = K_d^{ref}\frac{(C - R_{ref2})R_{ref}}{(C - R_{ref})R_{ref2}} \Leftrightarrow C = \frac{(K_d^{ref2} - K_d^{ref})R_{ref}R_{ref2}}{K_d^{ref2}R_{ref2} - K_d^{ref}R_{ref}}$$

2.4.8 Heat Shock Element model

Our probabilistic model for HSEs was designed to capture interactions among the binding preferences of the three monomers that form the HSF homotrimer. The model consists of three PSSM-based submodels corresponding to the three 5 bp sequences (pentamers) that are bound by the HSF monomers. Each of these submodels is defined by two PSSMs, one "strict" and one "relaxed". These three submodels allow for eight possible combinations of strict and relaxed pentamer binding. Within each pentamer the positions are considered independent, as in standard PSSM models.

Formally, let a candidate 15 bp HSE sequence X_k be composed of random variables $X_{i,jk}$ where i is the pentamer index and j is the base position within that pentamer. Additionally, let each sequence have an associated unobserved random variable Y_k which indicates which of the eight combinations of strict/relaxed distributions are applied the corresponding $X_{i,jk}$ (Figure A.12). For simplicity, our model definition assumes that the middle monomer sequence has been reverse complemented and is therefore in the same orientation as the outer monomer binding sequences. We considered two versions of the model: a sparsely parameterized "constrained" version and a more parameter-rich "expanded" version, as described below.

Constrained version

This version of the model assumes that the three monomers share the same strict and the same relaxed PSSM-based sequence distributions. In addition, it assumes that the relaxed PSSM is defined as a more degenerate version of the

strict PSSM. This is accomplished by means of a single "degeneracy" parameter, which "pulls" the nucleotide distribution at each position toward the uniform distribution. Specifically, the nucleotide distribution at position j of pentamer i is defined as:

$$P(X_{i,j}^k = b|Y_i^k) = \begin{cases} f_j^b & \text{if } Y_i^k = strict \\ \frac{f_j^b+B}{1+4B} & \text{if } Y_i^k = relaxed. \end{cases}$$

where f_j^b is the probability of observing nucleotide b at position j of the monomer and B is the free parameter controlling how close to an uniform distribution the relaxed version is.

To estimate the model parameters from the HSE sequence data, we first held B fixed and then estimated the nucleotide frequencies and the prior probabilities of each strict/relaxed monomer combinations through Expectation Maximization (EM). A grid search was then used to find the value of B that maximized the model likelihood.

Estimating the model parameter updates for EM is simple for the prior but slightly more complicated for the nucleotide frequencies due to the interdependency between the strict and relaxed distributions. Nevertheless, it can be solved by using a Lagrange multiplier together with the derivatives of the expected complete log-likelihood. This produces an estimator that depends on the Lagrange multiplier and requires the use of a root finding method as part of the maximization step of EM. Figure A.13C shows the results of the parameter estimation. To initialize the optimization procedure, the nucleotide frequencies were estimated from the high-confidence in vivo HSEs from [53].

Simulation study

To test the performance of this model, we estimated unconstrained strict/relaxed matrices from real data and simulated data under various distributions of Y_k. All of the parameters of the simplified model were then estimated from this simulated data, and the posterior distribution of Y_k under the model was compared with the values used for simulation (Figure A.14).

Expanded version

This version of the model allows for completely separate PSSMs for the three pentamers, and completely separate strict and relaxed versions of each of these models. It has $5 \times 3 \times 3 \times 2 = 90$ PSSM parameters plus seven free parameters for Y_k, for a total of 97 free parameters. Parameter estimation is again accomplished by expectation maximization, but in this case the parameter updates are trivial.

2.4.9 Chromatin effect and DNase I hypersensitivity models

The chromatin effect and DNase models are rule ensemble models, estimated using the RuleFit R package. This package was also used to estimate the relative importance of the model covariates. The covariates were obtained from modENCODE tracks, taking the mean value over a 200 bp window centered on the target point. Furthermore, these data were filtered to contain only points that had a value for every covariate used.

Chromatin effect model

This model estimates the ratio between the in vivo and in vitro intensities at each site from a set of chromatin covariates. Ratio values were obtained from the measured intensities using the in vitro HSE cluster coordinates. The set of HSE clusters was pre-filtered by finding a threshold on the in vitro intensities that approximately minimized the differences between experimental replicates. The threshold was obtained as follows: 1) for each candidate in vitro threshold value, collect the mean absolute difference between experimental replicates for the ratios computed using in vitro intensities above that value; 2) compute the mean of the values collected in the previous step; 3) pick the first in vitro threshold that falls below the mean.

The model was estimated on the selected HSE clusters using different combinations of chromatin covariates. For each particular combination, an estimated Pearson correlation value was obtained from ten-fold cross validation. Furthermore, to obtain the figures presented in this paper, the data was split into 60% training data and 40% test data. The model obtained on the training data was used to make the test data predictions shown in the figures and the corresponding Pearson correlation.

DNase I hypersensitivity model

The data set used for this model was independent of the HSE clusters. 10K points were randomly sampled from across the genome with the restriction that the points did not fall within the regions shown in the figures presented in this paper or within 200 bp of the HSE cluster sites. These 10K points were used as

a training set to build the model used to make the predictions for the browser tracks and at the HSE cluster regions. They were also used to estimate the Pearson correlation via ten-fold cross validation.

Prediction tracks

To produce the in vivo intensity prediction tracks, the chromatin model was applied to a version of the in vitro intensities that were scaled so that they would be comparable to the in vitro intensities. To obtain the scaling factor, we selected the top seven most accessible isolated HSE clusters (as measured by DNase hypersensitivity) that had a significant read count. The reason for these restrictions is that highly accessible sites should be good proxies for sites that are not being influenced by chromatin effects and the sites with significant in vitro intensity should produce better estimates of the in vivo to in vitro ratio used for scaling (see Figure A.4 for point choices).

The browser tracks were produced by collecting values in 50 bp steps with a 100 bp window average and applying the respective model and scaling (if needed). Values were then smoothed with a Gaussian kernel having a 100 bp bandwidth.

CHAPTER 3

ANALYSIS OF TRANSCRIPTION START SITES FROM NASCENT RNA IDENTIFIES A UNIFIED ARCHITECTURE OF MAMMALIAN PROMOTERS AND ENHANCERS[1]

3.1 Introduction

Regulation of transcription is a critical process for directing cell fates during organismal development and is necessary to maintain homeostasis throughout the lifespan of all organisms. Promoters and enhancers are major control hubs for transcription that integrate information from a multitude of signaling pathways through binding of signal-responsive activators and repressors. Therefore, accurately mapping and characterizing these regulatory regions is essential for defining how cell-specific transcriptomes are generated and maintained.

In mammalian cells, transcription initiation occurs on both strands at promoters and enhancers alike [74][27][123][78]. While this "divergent" transcription initiation remains incompletely understood, it is nevertheless a characteristic signature that can be exploited in the identification of active regulatory elements [56][97][4]. The signature of divergent transcription is particularly evident when transcriptional activity is assayed using the Global nuclear Run-On sequencing (GRO-seq) method, owing to its high sensitivity for all transcriptionally-engaged RNA polymerase regardless of subsequent transcript turnover rates[27][55][97]. We have recently shown that sensitivity for detect-

[1]The content of this chapter is the result of a joint work with Leighton Core, (submitted for publication). Leighton Core developed the GRO-cap assay and, together with Collin Waters, conducted the experimental assays. I developed and conducted the statistical analysis. Writing and interpretation was a joint effort with the additional collaboration of Charles Danko.

ing transcription initiation can be further improved by enriching the nuclear run-on RNA pool for 5-7meGTP-capped RNAs, using a protocol called GRO-cap[80][81] (see Methods). When applied to the nematode, *C. elegans*, GRO-cap was able to identify previously unknown, rapidly-degraded, TSSs from trans-spliced genes, thus allowing identification and study of true gene promoters [80]. In this article, we apply GRO-cap to human cells and show that it efficiently and precisely maps TSSs of coding and non-coding RNAs regardless of the resulting stability of the transcript. Thus GRO-cap provides a more complete picture of genome-wide initiation than CAGE, which detects mainly stable transcript initiation. Using our comprehensive, GRO-cap-based annotations of TSSs, we then report a detailed analysis of transcription initiation sites that sheds new light on the architecture of both promoters and enhancers across the human genome.

3.2 Results

3.2.1 Identification of Transcription Start Sites in Human Cells using GRO-cap

We prepared GRO-cap and GRO-seq libraries from human lymphoblastoid B-cell (GM12878) and chronic myelogenous leukemic (K562) cell lines. We have also included a PRO-seq dataset in K562 cells sequenced to high depth (365 million reads; Table B.1). These are both "Tier 1" cell lines in the ENCODE project, allowing us to take advantage of abundant publicly available functional genomic data [25]. The GRO-cap assay efficiently captures TSS information from

nascent transcripts, evidenced by a dramatic enrichment of GRO-cap signal at gene promoters and enhancers (Figure 3.1a, b, Figure B.1a). Figure 3.1a shows a specific example of the classic globin locus where divergent transcription is seen from active regions including the epsilon globin gene and the upstream hypersensitive sites (HS) that mark enhancers [101].

To comprehensively identify TSS candidate sites using our data, we developed a hidden Markov model (HMM) that contrasts GRO-cap with data from control experiments, in which the critical CAP-removing enzyme, tobacco acid pyrophosphatase (TAP), is omitted (Figure B.2a,b, Methods). In total, 120K TSSs were identified in each cell line (117,613 for GM12878, 128,471 for K562), within the range previously reported (80 to 150 K)[78], [138][143][33]. Predicted TSS regions are narrow (mean 57 bp, with 95% under 140 bp), but include 69% of all GRO-cap TAP+ reads. Our candidate TSS regions capture both those with sharp TSSs as well as those with a more dispersed signal (Figure B.2c,d; Methods). Predicted TSS regions overlap well with predicted regulatory regions from the same cell type, as 93% are contained within predicted enhancer or promoter regions based on patterns of histone modifications (ChromHMM regions)[37]. However, our predictions have much higher resolution than ChromHMM regions, as they cover a much smaller fraction of the genome, 1.6% in overlapping ChromHMM versus 0.5% in our TSS set (total ChromHMM covers 6%), (Figure B.2e,f).

In comparison to CAGE, GRO-cap shows a similar composite profile when aligned to annotated gene TSSs (Figure 3.2a). Importantly, GRO-cap appears to have reduced "background" levels in genes, as fewer reads map to introns and exons compared to CAGE (Figure 3.2a,b), resulting in more GRO-cap

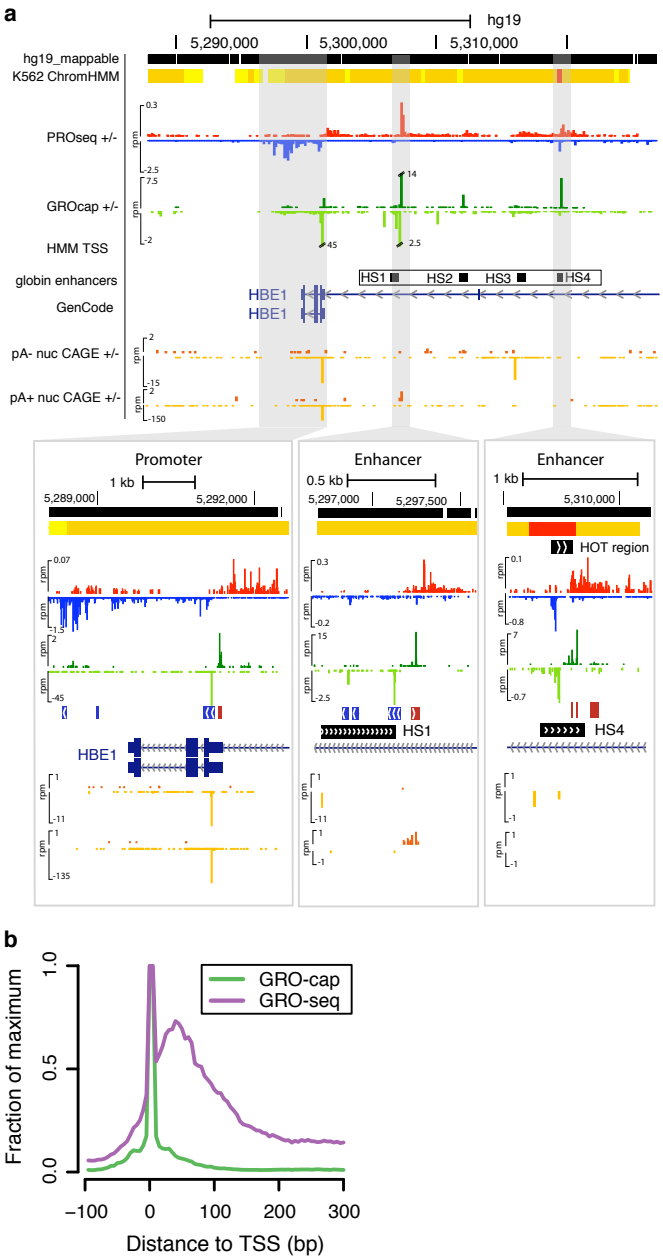

Figure 3.1: GRO-cap identifies TSSs in promoters and enhancers (a) A UCSC genome browser[75] shot of the globin locus near the LCR using K562 cell line data sets generated or used in this study. The locus contains a portion of the beta-globin locus, including the globin epsilon gene and LCR enhancers. The insets are zoomed in views of the shaded regions that show the divergent GRO-cap (+ strand: dark green, - strand: light green) signal at the epsilon-globin promoter (left) and two enhancers associated with the hypersensitive site (HS) 1 (center) and HS4 (right). The locations of the HS sites are taken from probe location in [5]. ChromHMM regions track is shown on top, with predicted promoters indicated in red and enhancers in orange. Note that CAGE signal (+ strand: dark orange, - strand: light orange) is at background levels in the enhancer region. (b) GRO-cap dramatically enriches the signal for initiation sites when compared with GROseq. Composite GRO-seq and GRO-cap reads from the cell line plotted relative to the TSS.

reads landing in promoter or enhancer regions (Figure B.3a). The decreased background for GRO-cap possibly results from differences in the methodologies (cap-trapping [126] versus the oligo-capping method [94] used to capture capped transcripts. Furthermore, selection for nascent RNAs avoids overrepresentation of post-transcriptionally capped RNAs that can accumulate in cells [38].

A major strength of the GRO-cap method is the ability to detect initiation of rapidly degraded noncoding RNAs (ncRNAs)[80]. For instance, GRO-cap can readily detect TSSs from upstream antisense RNAs (uaRNAs) at protein-coding promoters, whereas this signal is often absent in CAGE data (Figure 3.1a, 3.2c). Additionally, GRO-cap coverage of enhancer regions, predicted from histone modification patterns [37], is improved over CAGE (Figure 3.1a, Figure 3.2d,e, Figure B.3b,c). We further characterized enhancers captured by GRO-cap by contrasting our TSSs with ChromHMM enhancers and open chromatin (DNase hypersensitive (DHS)) regions. This approach subdivides ChromHMM enhancers into three main classes: poised (ChromHMM only); open (ChromHMM and DNAse HS); and transcribed (ChromHMM, DNAse HS and GROcap TSS) (Figure 3.2f). The transcribed subset is enriched for signs of positive regulatory activity, namely increased evidence for TF binding (wellington footprints[105]; Figure 3.2g), distal chromatin interactions (ChIA-pet [84]; Figure 3.2h), and reduced CpG methylation [96] (Figure 3.2i). In addition, the various histone modifications change in expected patterns among poised, open, and transcribed enhancers (Figure B.4). Together with our previous work [80], these results show that GRO-cap efficiently maps TSSs at active regulatory regions. Notably, we both efficiently capture regions identified by CAGE and enrich for active enhancer regions missed by CAGE due to the instability of enhancer RNAs (eR-

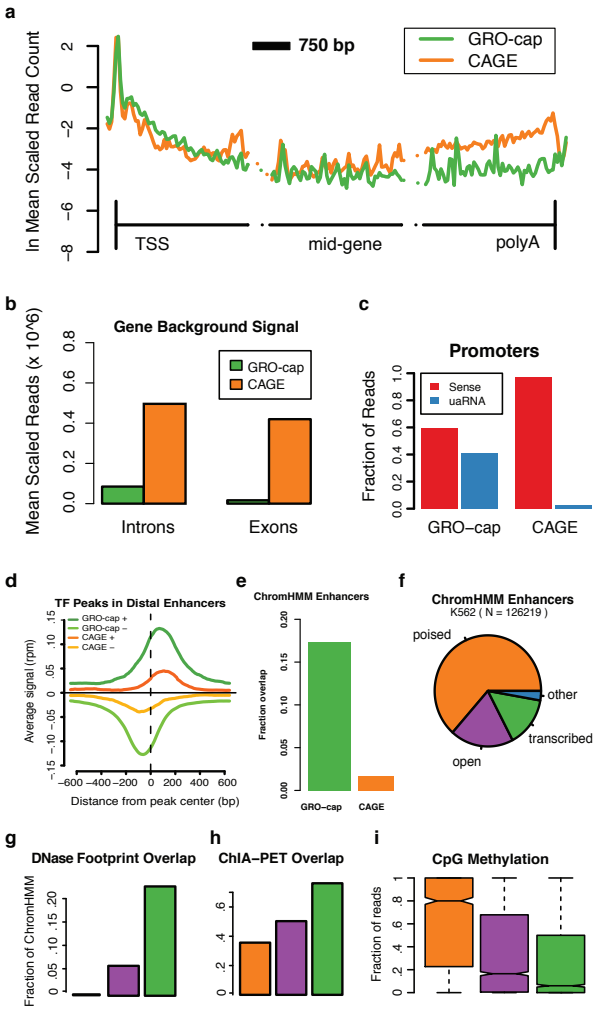

Figure 3.2: Comparison of PRO-cap with CAGE (a) GRO-cap and CAGE profiles at protein-coding genes. Genes are broken into three 3 Kb regions covering region around the TSS, the middle of the gene, and near 3-cleavage/poly-A site. The vertical lines represent the TSS and 3-cleavage site. (b) Average read density in interior introns and exons (excluding the first and last of each) as a measure of GRO-cap and CAGE background signals. (c) GRO-cap and CAGE relative fraction of reads aligned to sense and divergent (uaRNA) directions at protein-coding genes (counted within underlying ChromHMM region). (d) GRO-cap and CAGE profiles at transcription factor peaks in distal enhancers. (e) Fraction of ChromHMM regions containing a detectable GRO-cap (green) or CAGE (orange) TSS. (f) Comparing enhancer regions based on chromatin marks (ChromHMM Enhancers, Ernst. et al) with DNAse HS (OpenChrommatin consortium) and GRO-cap, reveals three main classes of enhancer regions, poised (no DNAse HS peak nor GRO-cap TSS; orange), open (DNAse HS peak, but no GRO-cap TSS; purple) and transcribed (DNAse HS peak and GRO-cap TSS; green), and a negligible 'other' (no DNAse HS peak but with GRO-cap TSS; blue). (g-i) These three classes represent a progression in terms of functional activity, as measured by (g) an increase in detectable TF footprints (Wellington footprints on DNAse HS,), (h) chromatin links (ChIA-PET overlap,) and (i) a significant reduction in CpG methylation between each transition.

NAs).

3.2.2 "Stable" and "Unstable" RNAs at Transcription Start Sites

Transcription at promoters and enhancers initiates in both forward and reverse directions, as seen with GRO-seq in previous work [27][56]. To simplify downstream analysis, we created a set of "divergent TSS pairs" that was filtered against cases of partially overlapping initiation pairs (Methods). The resulting set is composed of 22,443 TSS pairs (GM12878, 38% of all predictions; K562, 24,894 pairs or 39% of predictions). As both cell lines show similar results, we will refer to GM12878 data unless otherwise stated. We then classified GRO-cap-based TSSs into those giving rise to "stable" transcripts (captured by CAGE and GRO-cap) and those that produce "unstable" transcripts (captured only by GRO-cap) (Figure 3.3a). In practice, our TSS regions were classified as unstable in the absence of CAGE reads and as stable if they contained at least 8 CAGE reads (Methods). This threshold is conservative, and above the estimated CAGE background in introns (Figure 3.3b; gray bars). We focused on high-confidence sets of both stable and unstable transcripts by further requiring a high GRO-cap signal (minimum number of reads above the 20% quantile). The distinction between stable and unstable transcripts is also apparent from other RNA-based assays. For instance, stable TSSs have strong RNA-seq profiles (Figure B.5), whereas unstable TSSs have very weak or non-existent RNA-seq profiles. These patterns hold for both the polyA-plus and polyA-minus versions of CAGE and RNA-seq, indicating this difference is not simply due to a change in

poly-adenylation.

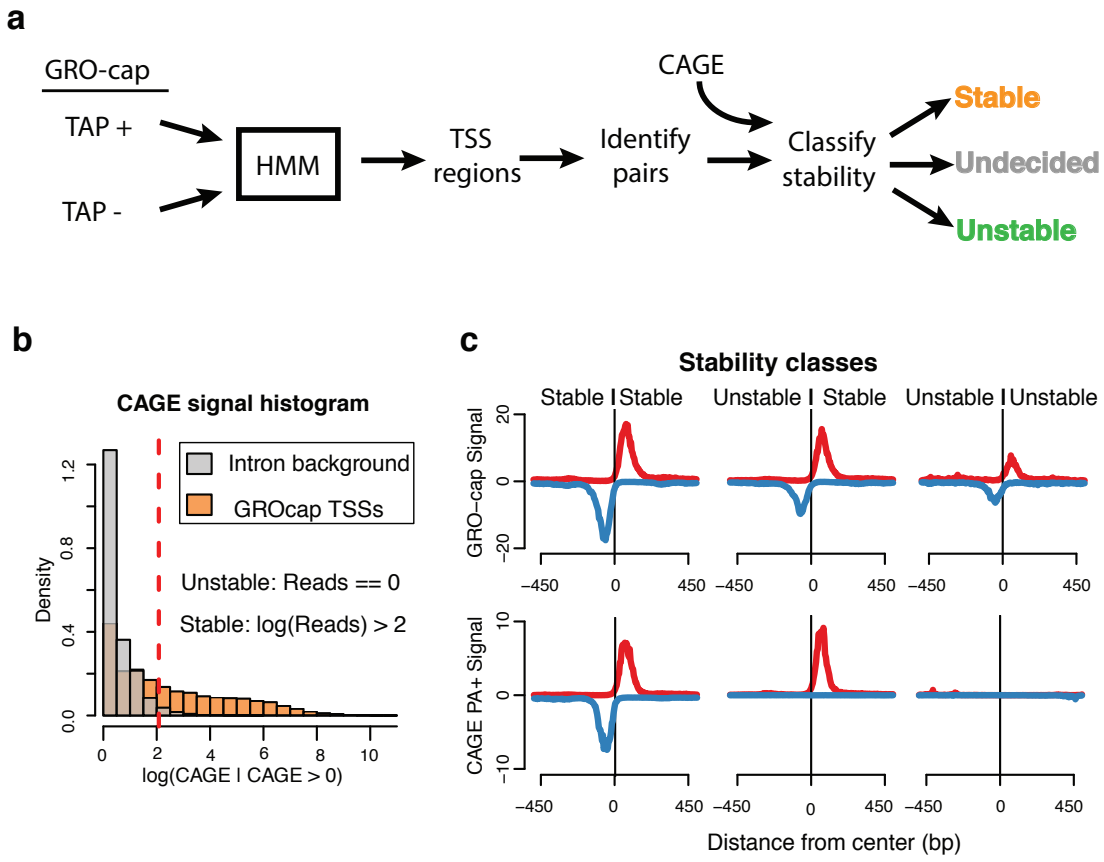

Figure 3.3: TSS identification and classification. (a) TSS regions were identified with a hidden Markov model (HMM) from GRO-cap reads and control, and combined into pairs of divergent TSSs which where then classified according to the presence of CAGE signal. (b) CAGE signal histogram at GRO-cap TSSs (orange) overlaid with CAGE background signal estimated from introns (grey). TSSs were classified as stable if above threshold indicated by dashed red line, or unstable if they contain no CAGE reads. (c) Composite profiles of GRO-cap and CAGE aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable (left), unstable::stable (center), unstable::unstable (right).

Divergent TSS pairs fall in three stability classes: Stable:Stable (SS), Unstable:Stable (US) and Unstable:Unstable (UU) pairs, which were the basis for subsequent analysis (Figure 3.3a,c, Figure B.6a). These classes cover a wide range of directional transcription preferences irrespective of the stability class (Figure B.6c, Methods). The stability of individual TSSs and, by extension,

the classes of TSS pairs generally correspond to different transcript annotation types (Figure B.6b). Histone marks [128] show a distinct pattern depending on the pair stability class. For instance, the H3K79me2 elongation mark shows a clear association with the stable direction (Figure B.7). Furthermore, UU pairs are more strongly enriched for the proximal H3K4me1 signal than are SS pairs, while SS pairs are enriched for the H3K4me3 mark. Overall, the SS and US classes are enriched in chromatin signatures associated mainly with promoter regions, and correspond to various stable transcripts such as protein-coding genes and long intergenic non-coding RNAs (lincRNAs) (Figure B.6b). Although lincRNAs can be found paired with promoters and at enhancers [19], GENCODE [57] lincRNAs are largely stable by our classification. Altogether, our TSS pair classes show distinct patterns of RNA and chromatin marks and correctly cover the expected transcript annotation types. However, our data-driven classification condenses all transcript types into three fundamental types for further analysis in an annotation independent fashion.

3.2.3 Transcriptional Level Explains Major Differences in Histone Modifications Between Enhancers and Promoters

In general, the ChromHMM distinction between promoters and enhancers follows our TSS classes, with SS and US pairs mainly found at active promoters and UU pairs mainly found at enhancers (Figure 3.4a). However, we observe that a large fraction of UU pairs are classified by ChromHMM as active promoter regions. This observation is unexpected given that active gene promoters should produce a stable transcript in at least one direction. Inspection of the

UU pairs classified as active promoters revealed that they have stronger PRO-seq signals than UU pairs classified enhancers (Figure 3.4b). This raises the possibility that these UU pairs classified as promoters are instead highly active enhancers and may share similar histone mark characteristics to those at active promoters.

In order to investigate the relationship between transcription level and histone marks, we defined a set of stable TSSs from US pairs proximal to annotated protein-coding genes (putative promoters) and contrasted them with TSSs identified from UU pairs in TF ChIP-seq peaks that are distal from genes (putative enhancers). Although these promoters are generally more highly transcribed than the enhancers (see Discussion), we observed that the H3k4me3/H3k4me1 ratio at both the promoters and enhancers scales with the corresponding level of Pol II (Figure 3.4c, d). Expanding this analysis to all GRO-cap-identified TSSs in our TSS pairs, we observed that transcription-associated histone modifications are directly related to the transcription level and this relationship is maintained independently of transcript stability (Figure 3.4e). That is, as the level of transcriptionally-engaged Pol II increases at TSS pairs, so do transcription-associated histone modifications. The CpG-binding protein, Cfp1, has been implicated in transcription-independent deposition of H3K4me3 through its recruitment of Setd1[24]; however, its DNA binding domain is dispensable for targeting H3K4me3 to active genes. Our observed increased H3K4me3 at UU pairs of active enhancers is likely to be dependent on transcription rather than simply on an increase in CpG content for this subset of enhancers, because we see a clear difference in the CpG content at stable and unstable transcripts and nascent transcription at both (Figure B.8). Thus, the difference in histone modifications at promoters and enhancers is not specific to the type of regulatory

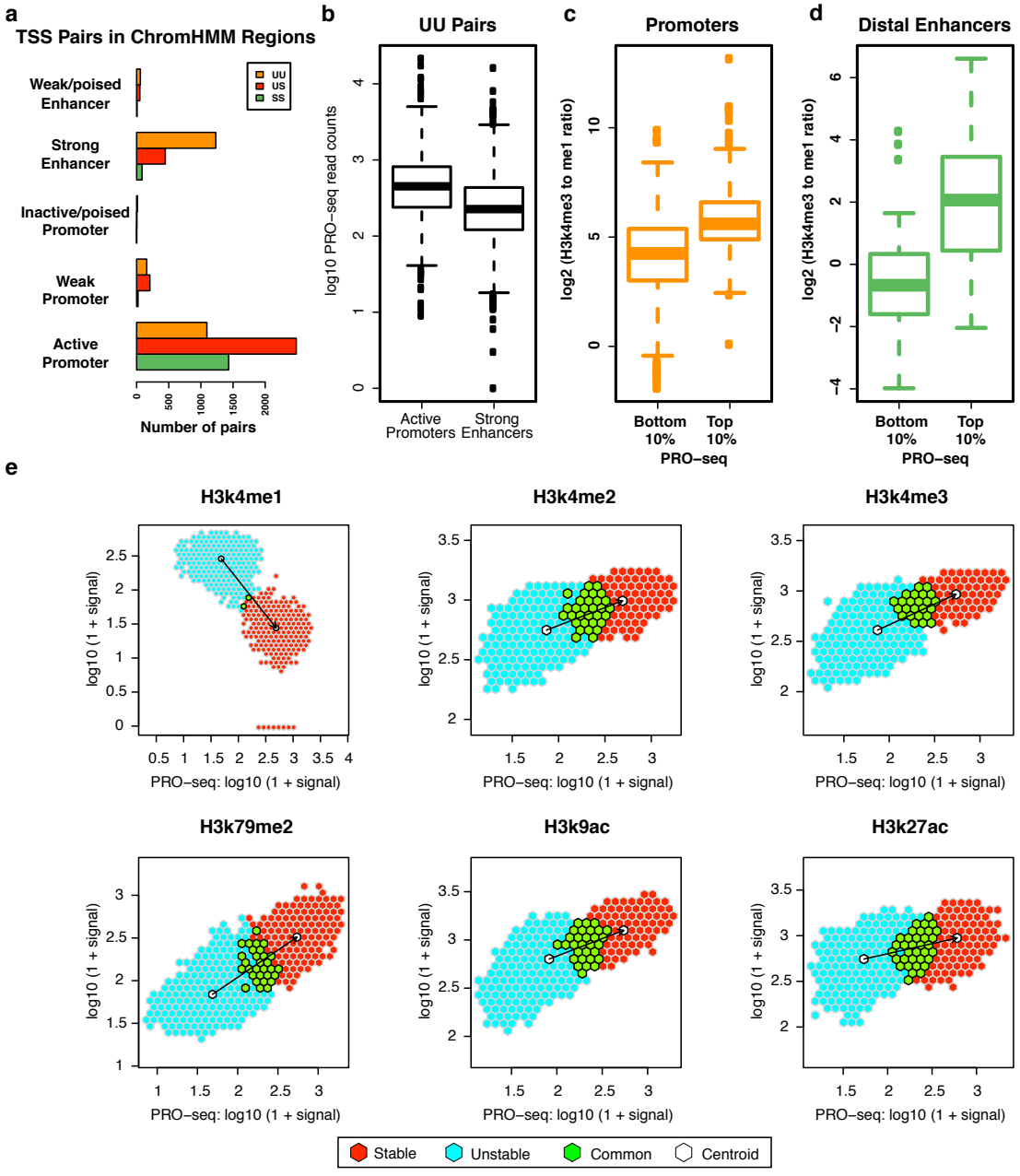

Figure 3.4: Histone marks at enhancers and promoters scale with Pol II intensity. (a) Number of TSS pairs from each stability class mapping to different regulatory regions as designated by ChromHMM. (b) UU pairs mapping to active promoter regions have a higher PRO-seq signal than those mapping to strong enhancer regions, where active promoters and strong enhancers are defined by ChromHMM. (c-d) Ratio of mono to tri methylation of H3k4 at top and bottom deciles of PRO-seq signal in both (c) promoter and (d) enhancer TSS regions. (e) PRO-seq signal versus indicated histone modifications at TSS regions. Signal is further split between TSSs classified as unstable (light blue), stable (red), and points that overlap between the two (light green). Centroid for each subset in white.

element or transcript, but rather, this difference appears to be more fundamentally associated with the level of transcription.

3.2.4 Transcription Factor Binding Appears to Drive Initiation Architecture

Having found an initial link between histone modification patterns and transcription levels, we turned to other features of initiation regions that might determine transcript outcome. We start by more closely examining the architecture of TSS regions. We previously reported that divergent initiation at human promoters is typically separated by 250 base pairs[27]. However, our original assay was not designed to detect TSSs per se. Here, we use our high confidence TSS pairs to refine this analysis and show that divergent initiation occurs, on average, 110bp apart (Figure 3.5a) with relative small variations between TSS pair classes (Figure B.9). While divergent initiation is less common in *C. elegans*, our estimates of distance between divergent pairs in that species is nearly identical [80]. Despite the narrow distance, high-resolution ChIP-exo [138] of two general transcription factors (GTFs) that bind core promoters (TBP and TFIIb) and Pol II reveals an independent transcription initiation complex forms in each direction at divergent TSS pairs at promoters and enhancers (Figure 3.5b).

Transcription initiation is often closely followed by promoter-proximal pausing. ChIP-exo data has revealed that the majority of Pol II at promoters is likely to be in a paused state [138]. Thus, we hypothesized that there might be some interplay between the strength and location of pausing and divergent TSS distances. Although we observe distinct pause modes (proximal-focused

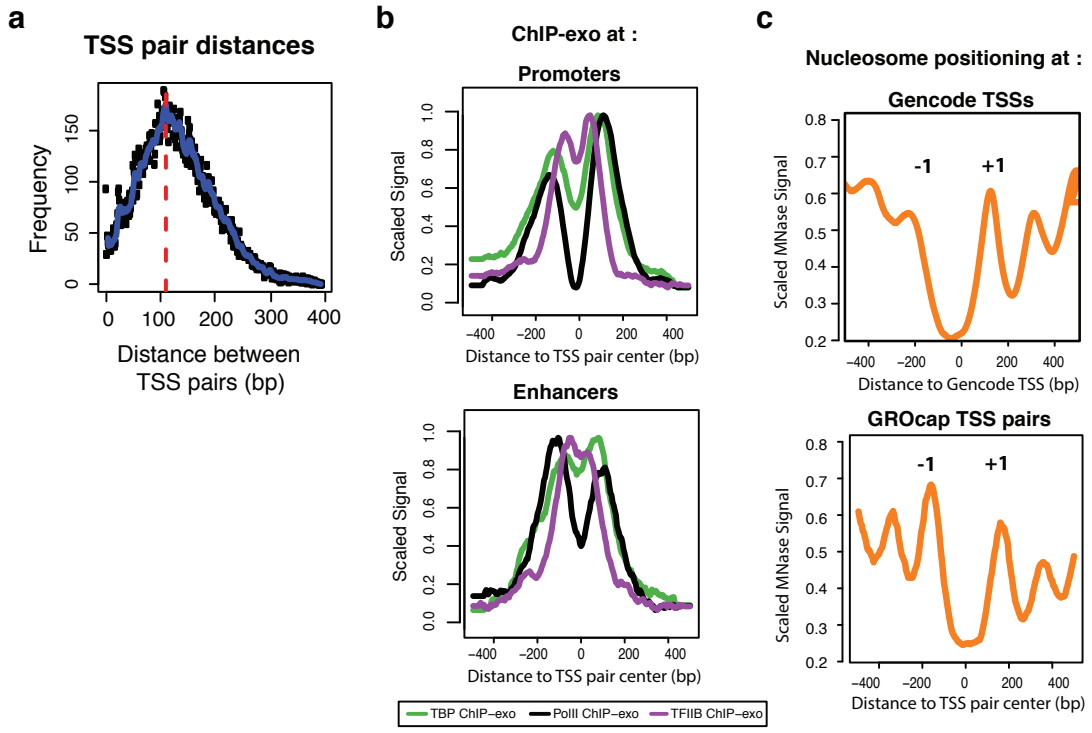

Figure 3.5: Architecture of TSS pairs. (a) Divergent TSSs are tightly packed, with an estimated 110 bp inter-TSS distance, as estimated from the overall distribution of opposing strand read distances. (b) ChIP-exo profiles [138] for Pol II (black), TBP (green) and TFIIB (purple), centered on TSS pairs and split between promoter (top) and enhancer (bottom) regions (ChromHMM). (c) Mnase-seq profiles at protein-coding promoters, aligned either by GENCODE annotations (top) or GROcap TSS pair centers (bottom). Peaks corresponding to -1 and +1 nucleosomes are indicated.

and distal-dispersed, as previously found in *Drosophila* [81], we find no effect of these modes on divergent initiation distances (Figure B.10a-c). We also observe no effect on peaks TFIIB positions with different pause modes (proximal-focused and distal-dispersed)(Figure B.11). This, along with the similar divergent TSS distance results from *C. elegans* (where pausing is rare), suggests that pausing location does not feed back and influence the locations of divergent TSSs.

Although we find symmetric initiation and GTF binding at divergent promoter TSSs, nucleosome positioning is thought to be asymmetric at promoters.

Typically, when aligned to GENCODE TSSs, there is a well-positioned downstream nucleosome (+1 nucleosome), whereas the upstream nucleosome (1 nucleosome) has more variable positioning [122] (Figure 3.5c, top). On the other hand, nucleosomes are reported to be strongly positioned at both sides of TF-bound enhancers [46] (Figure B.12). However, when aligned to the center of our TSS pairs, we clearly see that both nucleosomes flanking the protein-coding US and SS TSSs are well-positioned (Figure 3.5c, bottom), with similar profiles to those at enhancers. Thus, the symmetric architecture of initiation regions applies universally to promoters and enhancers.

The observed symmetries of nucleosome positioning and core promoter factor binding raise the issue of how sequence-specific TFs bind within this context. Using TF ChIP-seq data from ENCODE, we observed four main preferences for pair classes by TFs (Figure 3.6a, Figure B.13-B.26): factors that bind preferentially at SS pairs (e.g., GABP); factors that bind preferentially at UU pairs (e.g., PU1); factors that bind indiscriminately at all pair classes (e.g., BCL3) and factors with a preference for US pairs (e.g., CTCF). In addition, we observed two clusters by relative position of binding sites within divergent TSS pairs (Figure 3.6b,c): central binding factors (e.g., SP1) and TSS-proximal binding factors (e.g., YY1). We are limited by the ChIP-seq sets available, but given the datasets used (N = 84), most factors fall into the central binding cluster (binding profile peaks in the center between divergent TSSs; N = 73) versus the TSS binding cluster (binding profile peaks over TSS position; N = 10) (Table B.2). Interestingly, the TSS-proximal binding cluster includes both GTFs such as TAF1 and transcription repression factors such as NRSF and Pml (Figure 3.6d), suggesting a potential involvement in transcript stability determination or preferential targeting of regulation to stable transcripts. These results evoke a model where

activation of initiation and the surrounding chromatin architecture are driven by a central-binding activator at both promoters and enhancers alike, and modulated by TSS-proximal TF binding.

3.2.5 Sequence Predictors of Transcript Stability

Because DNA sequence is known to influence initiation, productive transcription, RNA processing and stability, we also examined the sequence composition near our TSS pairs. In general, we find that sequence conservation and nucleotide frequency are indicative of transcript stability (Figure B.27a-c). In particular, SS TSSs are associated with increased C and G nucleotides and increased CpG di-nucleotides within and around pairs. In contrast, UU TSS pairs are depleted for C, G, and CpG. US TSS pairs display a combination of these two patterns. Despite these biases, we see similar frequencies of core promoter elements (TATA and Inr) in the expected positions at all classes of TSS pairs (Figure B.28a,b). This observation is consistent with ChIP-exo detection of GTFs at all classes of TSS pairs (Figure B.28c), indicating that other mechanisms might be dictating the production of stable versus unstable transcripts. Indeed, recent work has shown that sequences that direct the binding and activity of poly-A dependent termination machinery or the U1 splicing complex work antagonistically to direct unstable or stable transcription, respectively, at protein-coding genes [3][100]. In this model, 5-splice sites (SS5) that bind U1 can suppress poly(A) site (PAS)-dependent termination, thus promoting productive elongation of protein-coding mRNAs.

To determine if there is a direct relationship between our transcript stability

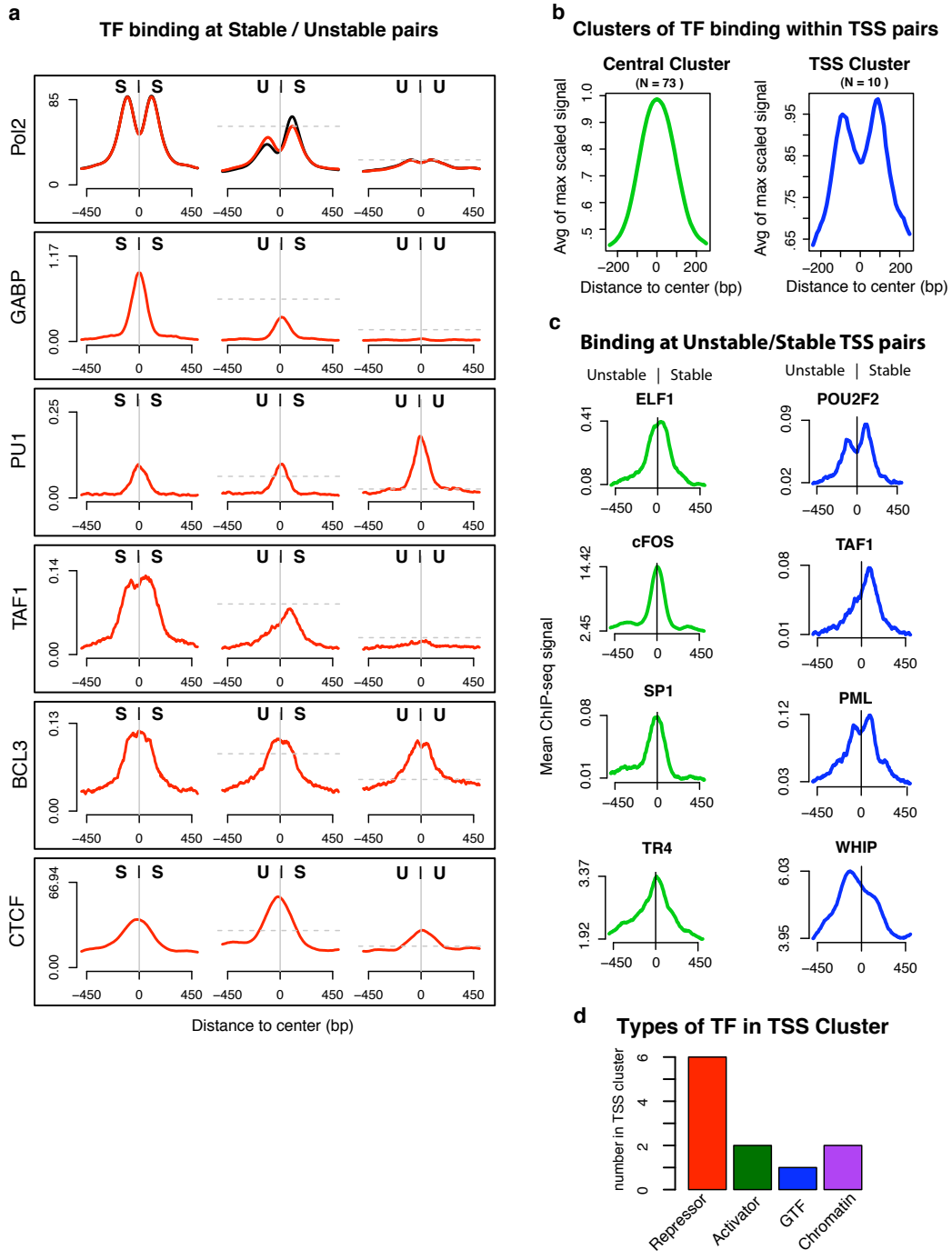

Figure 3.6: Modes of TF binding at TSS pairs. (a) Representative ChIP-seq profiles of different modes of transcription factor binding at different TSS pair stability classes. Signals are subject to paired subsampling to correct for Pol II signal dependency (top plot, Methods), (b) ENCODE TF ChIP-seq profiles, anchored on TSS pairs, cluster into two distinct groups, central binders (green) and TSS binders (blue). (c) Examples of the two positional modes of binding at US (Unstable,Stable) pairs. (d) Classification of factors within the TSS binding cluster. The total number of factors in d are greater than the number of TSS binding factors because factors can be part of more than one functional group (see Table B.2).

classes and the premature PAS-dependent termination, we scanned the regions downstream of TSSs for matches to the poly-A and SS5 motifs and observe that our stable and unstable TSS classes follow a pattern consistent with these reports (Figure B.29a,b). That is, the SS5 motif is enriched downstream of stable transcripts but depleted at unstable transcripts, and vice-versa for the PAS motif. We devised an HMM that incorporates SS5 and PAS motif models and used it to compare the likelihoods of SS5 binding sites before and after a polyA site (Figure 3.7a, Figure B.29c). Our results indicate that that SS5 binding sites strongly tend to precede the PAS on stable transcripts but not on unstable transcripts (Figure 3.7b). These results are consistent with previous observations for protein-coding genes, and importantly, they demonstrate that these sequence predictors of elongation hold for all TSSs, including those at enhancers. Furthermore, our HMM can be used to predict transcript stability to high accuracy (63%), suggesting that these motifs and their spatial relationship are strong determinants in this process.

The link between splicing and stability raises the question of what mechanism prevails at single-exon genes. To address this question, we applied our HMM to a set of expressed, single-exon genes (N = 105). We find that the most likely scenario for single-exon genes is the absence of both PAS and SS5 sites (Figure B.29d), suggesting that promoter-proximal depletion of PAS plays an important role in determining transcript stability even in the absence of splicing. Also, we compared CAGE to GRO-cap at single exon genes and observe a reduced ratio in comparison to length-matched spliced genes (Figure B.29e) indicating that the presence of the splicing machinery further augments transcript stabilization beyond the absence of PAS sites.

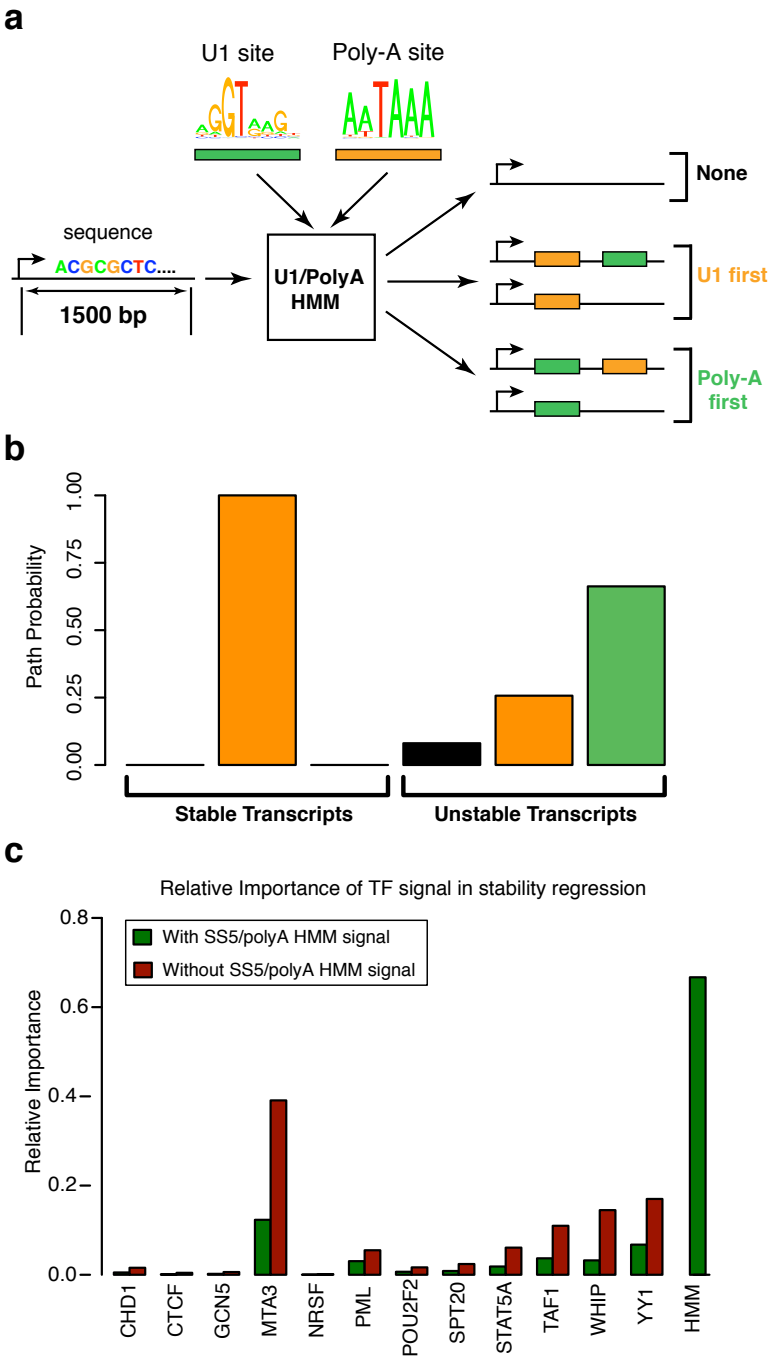

Figure 3.7: Determinants of RNA stability for both promoters and enhancers (a) Diagram of transcript U1/polyA classification. Each transcript (first 1.5kbp) is processed through an HMM to determine relative order and occurrence of SS5 and PAS elements. (b) Estimated path probabilities of alternative element occurrences (neither SS5 nor PAS: black, SS5 first: orange, PAS first: green) obtained by applying the EM algorithm to each transcript subset (stable and unstable TSS stability classes). (c) Relative importance of various transcript factors in a logistic regression of the stability classes, with (green) and without (red) including the U1/polyA HMM derived signal (posterior path probability of being in unstable class).

Finally, we used logistic regression to assess the relevance of TFs in the TSS-binding cluster to transcription stability. TFs, by themselves, explain only a small fraction of the variance in stability (R2 = 0.05). Furthermore, when the signal from the polyA/U1 HMM is also considered, their relative importance drops considerably (Figure 3.7c). These observations suggest that most of the information about stability comes from the presence or absence of early poly A sites and U1 splicing signals, but they do not rule out the possibility that some of these TFs may be components of the splicing pathway or contribute to feedback between splicing and expression levels.

In total, this work suggests a very similar architecture exists across all transcription initiation regions. Furthermore, regulatory regions are more distinguishable by post-initiation events such as elongation and transcript stability.

3.3 Discussion

Several studies have documented divergent transcription at promoters and enhancers [27][78][56][143][127][123], however, the nature and organization of initiation sites, their underlying DNA elements, and their relationships with TF binding and nucleosome positions have yet to be reconciled. In this article, we show that assaying nascent RNAs dramatically increases sensitivity for enhancer detection compared with methods that map accumulated RNAs. By contrasting our GRO-cap data with CAGE data, we are able to classify TSS pairs based on the stability of the resulting transcripts. Unstable transcripts are those that are likely targeted for immediate degradation by the exosome, and thus are unable (or less likely) to be discovered in assays that detect accumulated RNAs,

such as CAGE. By contrast, stable transcripts are detectable in both nascent and accumulated RNA pools. These classifications allow us to work directly from genome-wide functional genomic assays without reliance on genomic annotations. By analyzing these annotation-free TSSs together with DNA sequences and functional genomic data, we are able to catalog the precise nature of the structure and chromatin content at initiation sites. We find that the divergent TSS pairs at both promoters and active enhancers: 1) have similar frequencies of canonical core promoter elements, 2) have distinct transcription complexes at each member of a pair, 3) are separated by 110bp on average, 4) are bound by a central transcription activator, 5) are flanked on both sides by positioned nucleosomes, and 6) have histone modifications typically associated with transcription initiation, present in proportion to the amount of transcription. These results suggest a unified model for the mechanisms that govern transcriptional initiation at both enhancers and promoters (Figure 3.8a).

3.3.1 Architecture of Initiation Sites

We show that divergent initiation occurs within a window of 90-120 bp, which is a surprisingly narrow interval considering that a PIC makes contacts up to 50bp upstream and downstream from the TSS [28]. The close proximity of divergent initiation events and the evidence for bound TFs between them suggest that multiple independent polymerase complexes may not simultaneously occupy the same promoter. One possible alternative is that one polymerase initiates first and then pauses downstream, allowing enough space for a second polymerase to initiate upstream and in the opposite direction. Consistent with this hypothesis, high-resolution ChIP-exo data suggests that the majority of Pol II

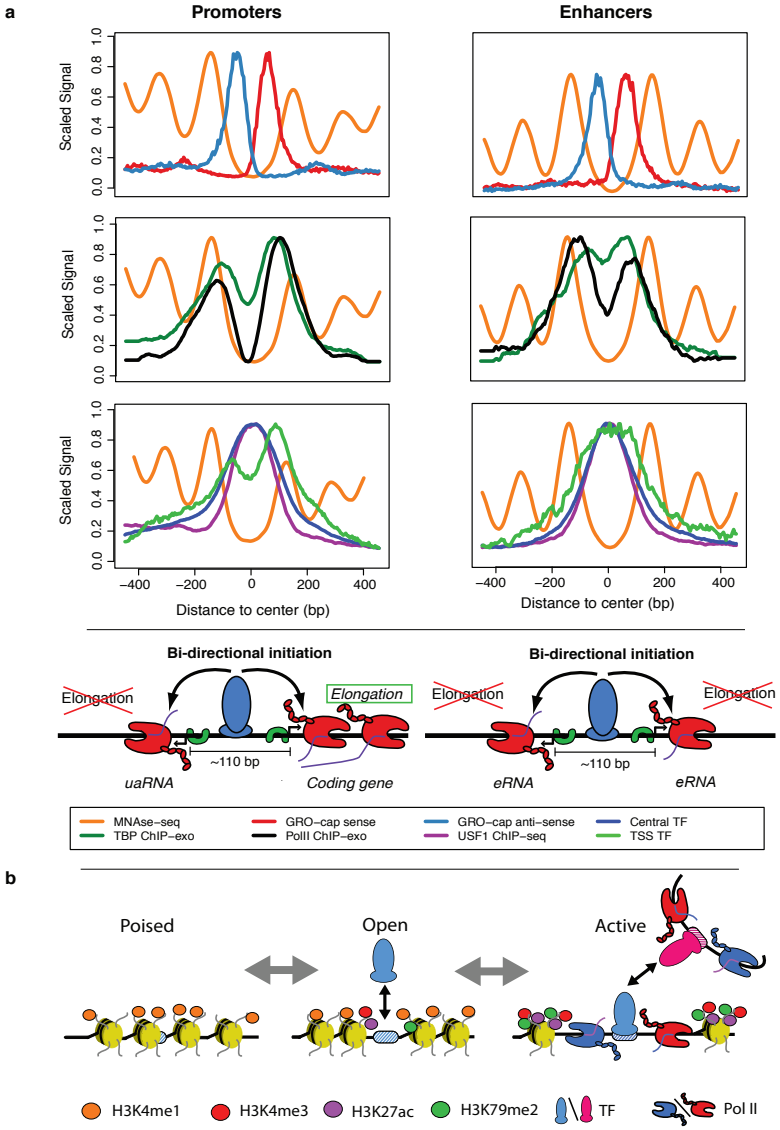

Figure 3.8: Unified model of transcription initiation at regulatory regions. (a) Our analysis of TSSs reveals a common structure across all initiation regions, including promoters and enhancers. In both cases, (first row) a tightly packed (110 bp apart) divergent TSS pair (+ strand: red, -strand: blue) surrounded by well-positioned nucleosomes (orange), with independent pre-initiation complexes (separate TBP (green) and Pol2 ChIP-exo peaks (black), second row) and sharing two distinct TF cluster binding modes (central: green, over TSS: blue; third row). We propose that central, activator TF binding (USF1 example: purple), in conjunction with core promoter elements, determines the positioning of the divergent initiation sites. Finally, DNA sequence properties (not depicted here), possibly in cooperation with other factors, determine the resulting transcript type (stable/elongating: protein coding, unstable/terminating: uaRNA, eRNA, etc.). (b) A model depicting possible progression of enhancer states from chromatin marked but largely inaccessible regions (left), followed by more open regions through TF binding (center) and finally, active transcription, which brings with it the associated chromatin marks (in particular, H3K79me2 and H3K27ac and increased methylation levels of H3K4; right).

on chromatin of human cells (K562) is paused approximately 50bp downstream of the initiation site [138].

In spite of the tight spacing available within divergent TSSs, we find evidence for positional modes for TF binding in these regions, though our data does not address whether they occur simultaneously. Our results show that most factors bind between the two divergent TSSs (central binders), suggesting that they play a role in activation and are likely a major determinant of the overall architecture of initiation sites. On the other hand, the TSS-proximal transcription factors are primarily enriched for repressors, suggesting that certain repressors can act by preventing access of the transcription machinery to critical parts of the core promoter. The apparent tight spacing and organization of binding suggests that very few factors simultaneously bind at any given initiation region. This observation is in agreement with evidence for a small number of identifiable sequence motifs even when numerous factors are found in narrow regions by ChIP-seq [41]. Furthermore, ChIP-seq signals can reflect indirect binding (local or distal tethering), and coinciding signals for different factors can reflect events that occur within a population of cells, but not simultaneously within the same cell. Finally, the close relationship between TF binding and initiation in our model provides a possible explanation for why protein-coding coding genes typically have multiple associated mRNAs with small differences in TSS location (for example GENBANK mRNAs). These alternative TSSs likely result from the presence of multiple neighboring binding sites, for the same or different TFs, that compete as anchors for activator TFs. As a result, depending on cell type and condition, different TF binding events lead to small shifts of the initiation site.

Promoter regions are generally assumed to be quite broad, with promoter-associated TF binding sites spanning a multi-kilobase region near the TSS, but our results suggest that initiation regions are primarily defined by a relatively narrow 100-to-200 bp window. Part of this discrepancy in scale can be attributed to poor or incomplete annotation of genes, but it may also indicate that multiple independent initiation regions often act as neighboring enhancers. Although we have focused here on non-overlapping TSS pairs to simplify our analyses, we expect that overlapping TSS pairs will represent an aggregate of the local TF occupancies. In the future, it will be interesting to further investigate TF occupancy at these more complex regions with the help of higher-resolution assays, such as ChIP-exo [114].

While our focus here is on mammals, and in particular, human cell lines, features of our basic model can apply to other metazoans that show predominantly unidirectional transcription at promoters. For example, *Drosophila* shows little divergent transcription at promoters, but divergent TSS pairs at enhancers [26]. Because both promoters and active enhancers are transcribed in *Drosophila* [26][17][77], as in mammals, it seems likely that the absence of divergent transcription in promoters reflects constraints on Pol II orientation. For example, it is possible that transcription units that encode stable polyA transcripts in *Drosophila* have evolved sophisticated core promoters with multiple elements that dictate the orientation of Pol II [40]. Indeed, Kadonaga and colleagues [71] have demonstrated a rich collection of elements and interactions at the core promoters of *Drosophila* genes. This additional feature of the *Drosophila* genome appears to help distinguish promoters from enhancer in flies, but it is of limited use in making the distinction in mammals.

3.3.2 Transition between enhancer states

Our analyses of DNase-hypersensitivity and GRO-cap data at enhancers generally support the existence of, and potential progression through, at least three enhancer states: closed, open, and transcriptionally active. Previous work suggests that chromatin undergoes a progression from a closed chromatin state to an open state required for TF binding and activity at enhancers [111][20][147][29]. We envision that it is equally plausible to progress in either direction between states, thus, the poised state could represent enhancers that have yet to be activated, or dormant enhancers that are vestiges of past activity [130]. Interestingly, the closed and poised state resemble a form of pre-activated promoters recently observed during developmental transitions [142], providing yet another similarity between regulation at promoters and enhancers. In our model, the poised state could transition into the open and untranscribed state through binding of a pioneering transcription factor that is required to open the chromatin prior to full activation by another TF. Also, some poised enhancers could be open simply because they have relatively poor affinity for nucleosomes due to underlying sequences. Alternatively, permissive chromatin could arise concomitantly with TF binding and transcription [20]. In either case, the transition from the open or poised states to transcriptionally active state appears to result from strong binding of central, activating transcription factors (Figure 3.8b). The transcriptionally active enhancer sites identified here are almost certainly functionally active, as they are enriched for TF binding, active histone modifications, distal chromatin links, and depleted for repressive CpG methylation (Figure 3.8b). However, it will require further work to determine whether or not all functionally active enhancers (influencing the activity of target transcripts) generate local transcription. In any case, our results show that

distinct enhancer states can be detected with various combinations of genomic data, and that a distinguishing feature of many active enhancers is the presence of transcription.

3.3.3 Transcription level and histone modifications at enhancers and promoters

It is generally thought that distinct mechanisms selectively mark histones at enhancers and promoters [20]. However, we observe a strong positive correlation between absolute levels of transcription and histone modifications traditionally associated with transcription activity, including both marks of initiation (H3K4me3 and H3 acetylation), and early elongation (H3K79me2) [128]. In particular, enhancers are typically identified as having high levels of H3K4me1 relative to H3K4me3 [62][61], but we observe a strong positive correlation between absolute levels of transcription and the H3K4me3/H3K4me1 ratio at active enhancers, suggesting that differences in H3K4 methylation patterns at enhancers and promoters may simply reflect differences in transcription levels. Consistent with this observation, H3K4me3 has been detected at some active enhancers [103][79], and can be H3K4me3 can be deposited in a transcription-dependent manner. For instance, the COMPASS complex that contains the Set1 methyltransferase, interacts with the post-initiation forms of Pol II [125] that have also been shown to be present at active enhancers [79]. Why, then, are enhancers generally observed to have less transcription initiation and hence H3K4me3 than promoters if transcription factors work identically at both types of initiation sites? Several feedback mechanisms whereby elongation of transcription

positively contributes to subsequent rounds of initiation have been observed. One such example involves feedback from the elongation-dependent H3K4me3 mark that can interact with the GTF, TAF3. It seems unlikely that this solely explains the differences between initiation levels at promoters and enhancers since this mark is present at both locations. Based on our results that describe a general splicing-dependent difference in transcript stability at promoters and enhancers, a more plausible explanation would be feedback from the splicing machinery. Indeed, increased recruitment of GTFs to promoters in the presence of a U1 splice site has been observed [30]. In addition, the GTF, TAF15, has been shown to interact with the U1 snRNP providing another link between the splicing and initiation complex. In support of this, our results show that another TAF (TAF1) binds preferentially to the stable side of promoters. Therefore, mechanisms that underlie deposition of histone modifications at promoters and enhancers may be more related to elongation of transcription and associated feedback on initiation rather than selective targeting of these regulatory elements.

3.3.4 Definition, form and function of enhancers

The original definition of an enhancer describes a genomic locus that stimulates transcription of another locus independently of its position and orientation relative to the transcribed locus [8]. Enhancers are often defined as wide regions, spanning multiple kb, where an abundance of ChIP-seq peaks are found. In contrast, our observations reveal a short common initiation structure for active promoter and enhancer sites anchored on a central activating TF, suggesting that these wide regions may actually be collections of individual initiation units. Our

results also show that the eventual outcome of transcription initiation is largely determined by the surrounding sequence and, in particular, by the presence of splicing signals and absence of poly(A) signals. Consistent with these observations, spliced lincRNAs are stable transcripts found both in isolation and paired with protein-coding genes. Furthermore, the common distinction of relative methylation levels for H3K4 is shown to be associated with transcription level rather than being a defining characteristic of promoters or enhancers. What then is a proper description of an enhancer? 3D chromatin links bridging different initiation regions have been observed both between traditional enhancers and promoters and between pairs of promoters [84]. Thus, the implication is that any initiation region can function as an enhancer, through the central binding activator, irrespective of the fate of the local transcripts that are generated. Conversely, it is currently not clear whether some TFs can enhance distal transcription activity without generating local transcription. Precise, high resolution annotation of TF binding sites should help the field progress towards a better understanding of local and distal interactions and insulator effects.

3.3.5 Evolutionary implications

Our observations have implications for an intriguing potential relationship between divergent transcription and the origin of new genes. It has recently been shown that asymmetries in productive transcriptional elongation favoring the sense-coding direction at gene promoters can be explained by a disproportional tendency for promoter-proximal cleavage and polyadenylation shortly after initiation in the antisense direction, which appears to be associated with the frequent occurrence of PASs in upstream antisense regions of genes [3][100]. Fur-

thermore, PASs are depleted and U1 snRNP recognition sites (SS5s) are enriched in the sense direction, consistent with observations that the U1 snRNP complex protects pre-mRNAs from cleavage and polyadenylation [72][13]. Building on these observations, Wu and Sharp recently proposed a model for the evolutionary origin of new genes whereby short, unstable upstream antisense RNAs (uaRNAs) gradually increase in length and stability as mutations eliminate PASs and create new SS5s [146]. In this way, uaRNAs could develop, in a stepwise fashion, first into noncoding RNAs and then into protein-coding mRNAs, perhaps acquiring splicing capabilities along the way (which, in turn, would further improve stability). This process could be encouraged by positive feedback with transcription-associated mutational asymmetries, which are biased toward G and T nucleotides [50] and therefore would favor the formation of SS5s and the abolishment of PASs.

Although divergent transcription at enhancers has been acknowledged [146], most discussion has focused on the emergence of new lincRNAs or genes immediately upstream of existing genes. In this article, we have shown that transcription initiation occurs in a bidirectional fashion at thousands of enhancers that have fundamentally the same architecture of initiation as traditional promoters. Thus, if uaRNAs do indeed sometimes develop into genes, then the genome is replete with potential new genes, many of them far from existing genes. These observations suggest a possible "life cycle" for new protein-coding genes in which genomic sequences may first acquire regulatory function by chance mutations, then become transcribed through enhancer activity, then gradually produce longer and more stable transcripts, and eventually become translated and obtain useful functions at the protein level. It is conceivable that many lincRNAs have limited direct functionality and are intermediate steps or

discarded by-products of this stepwise process. Additional studies of nascent RNAs across cell types and species may help to shed light on these important evolutionary questions.

3.4 Methods

3.4.1 Preparation of GRO-cap, PRO-seq and GRO-seq libraries

GRO-cap libraries for K562 and GM12878 cells were produced precisely as described in Kruesi et. al [80]. 1×10^7 nuclei were used for each GRO-cap library or control. GRO-seq libraries for K562 cells were produced as described in [144]. GM12878 GRO-seq data is published in Wang et al [144]. PRO-seq libraries were produced as described previously [81], using the TruSeq^{TM} small RNA adapters (Illumina), and 5×10^6 nuclei.

3.4.2 Mapping of sequencing data

After sequencing GRO-seq and GRO-cap reads were trimmed to 30 bases, and mapped first to a single copy of the rDNA locus to remove related transcribed sequences. Reads that did not map to the rDNA were then mapped to the hg19 version of the human genome. Reads were required to be unique and have no more than two mismatches. PRO-seq reads (100 bases) were processed essentially as in Kwak et al. [81]. Adapters were removed with cutadapt [93], and then unique sequences 15bp or greater were that mapped to the hg19 genome were kept for further analysis.

3.4.3 Prediction of Transcription Start Sites

Pre-processing of GRO-cap Data

GRO-cap aligned data, normalized by total read counts, was summarized in fixed intervals of 10 bp along the reference genome, to increase the signal in low intensity initiation sites and "smooth" away minor misalignments between the TAP+ and TAP- conditions. Each 10 bp interval was assigned two values, one summarizing the TAP+ to TAP- signal differences and the other indicating the presence of a TAP+ "peak". To summarize the TAP+ to TAP- signal difference in each interval we assigned the interval to one of three categories: 1) "no signal" (TAP+ has zero reads); 2) "enriched" (TAP+ >TAP-); or 3) "depleted" (TAP- >TAP+ >0). To compute the binary "peak" indicator for an interval, we searched for "depleted" intervals (as per the above definition) within ten 10-bp intervals (100 bp) in either direction, and if at least two were found, we used their mean normalized read counts as an estimate of the local background level. The interval in question was then called "peaked" if its normalized read count was greater than twice the estimated local background level. We found that our final predictions were not very sensitive to the threshold for calling peaks, with a wide range of fold-enrichments producing numbers of predictions that differed by no more than 3%.

Design of the Hidden Markov Model

Previous CAGE studies have shown that TSS regions can be both "sharp" (highly peaked) and "broad" [21][121]. As such, we designed our hidden Markov model (HMM) to have a single background state (B) and two groups of

alternative states, representing non-peaked (M1) and peaked (M2) TSS regions (Figure B.2a). The M1 and M2 groups are each composed of three states, and within each group, these states share the same multinomial emission distribution for "no signal", "enriched", and "depleted" TAP+ read counts. In addition, the states have a conditionally independent emission distribution for the peak signal, set such that only the middle state of the M2 group permits "peaked" intervals. Because multiple peaks can occur in a single peaked TSS regions, the transitions among the states in the M2 group allow for zero or more steps between consecutive peaks (middle state). This design enforces a distinction between sharp and broad TSSs, while avoiding false positives due to highly local spikes in the data.

Parameter Estimation and Transcription Start Site Prediction

The free parameters of the model were set as follows. Most transition probabilities were set to zero or one according to the constraints of the model design (Figure B.2a), or were assigned values reflecting a non-informative uniform prior distribution over possible state transitions (for example, the transitions out of the first and last states of the M2 group). The two exceptions to this rule were the self-transition probabilities for the background state and the middle (peak-emitting) M2 state, which were assigned high (0.99) and low (0.1) values, respectively, because we expect peaks to be sparse along the genome. The emission parameters were set approximately based on empirical observations of TSS regions. In particular, we observed that background regions are mostly devoid of reads ($P(\text{“}no\ signal\text{''}) = 0.9$; $P(\text{“}enriched\text{''}) = P(\text{“}depleted\text{''}) = 0.05$). By contrast, non-peaked regions (M1 group; broad TSSs) are dense in "enriched" intervals

(P("*no signal*") = 0.09; P("*enriched*") = 0.9; P("*depleted*") = 0.01). Peaked regions (M2 group; peaked TSSs) have both "enriched" and "depleted" intervals, in varying proportions, but because this group is anchored by the "peaked" indicator, it is not sensitive to the exact emission probabilities as long as "no signal" is unlikely; therefore, for these states we used P("*no signal*") = 0.1; P("*enriched*") = 0.45; P("*depleted*") = 0.45.

TSS regions were obtained by running the Viterbi algorithm [139][110] on the pre-processed GRO-cap data, which finds the most likely path through the HMM given the data and the model parameters. The predicted TSS regions were then refined for further analysis as follows. First, regions of longer than 100 bp that were assigned to the M2 state group were split into constituent "peaked" subregions such that distances of at least 30 bp were maintained between them. In addition, all regions were trimmed of leading and trailing "depleted" (TAP- >TAP+) intervals. The effects of these post-processing steps can be seen in Figure B.2b.

3.4.4 TSS Paired Regions

A divergent TSS pair is composed of adjacent TSS regions in opposing orientations (a minus strand TSS region followed by plus strand TSS region) within 150 bp of each other (nearest edges). This threshold was set empirically, after manual observation of initiation sites, in order to capture the observed distances between divergent TSS regions (median nearest edge distance was 40 bp). TSS pairs were further filtered by requiring a high GRO-cap signal (minimum number of reads above the 20% quantile), so that we could reliably scale the various

signals of interest by expression level in downstream analysis.

3.4.5 Paired Subsampling

In our analysis of divergent initiation regions, we produced composite profiles for paired TSSs in a variety of ChIP-based assays. A challenge in interpreting these profiles is that the marginal distributions of transcription levels often differ significantly at members of each pair, and other signals of interest, such as ChIP-seq measures of TF binding, correlate strongly with transcription level. Thus, apparent differences in the signals of interest may simply reflect differences in overall transcription level. This is especially a problem for US pairs, because unstable TSSs tend to have substantially lower transcription levels than their stable counterparts.

To improve the interpretability of these plots, we generated composite profiles by a sub-sampling method that ensures the marginal Pol II ChIP-seq distributions are the same at the left and right TSSs. Briefly, we summarize each TSS pair by four values: the Pol II ChIP-seq values and the signal of interest, both at the left and right TSS. For convenience, the Pol II ChIP-seq values are discretized into bins. We then define a shared "target" distribution for Pol II by pooling the data for the left and right TSSs. Finally, we subsample from the collection of TSS pairs (summarized by their four values) in such a way that the left and right Pol II distributions exactly match the target distribution. This subsampling step is complicated by the dependency between the left and right Pol II distributions, but it can be addressed by a simple algorithm that performs a depth-first search over possible combinations of samples from the original distribution, branches

of which are terminated whenever the constraints on the subsample are violated. In practice, we also cull a tree if more than 1000 consecutive siblings of a search node are found to be invalid, under the assumption that we have reached an dead end in the search (exploring the full set of alternatives is too costly as the number can be exponentially large). The induced marginal distributions of values for the signal of interest at the left and right TSS are then compared. In this way, differences in the profiles that simply reflect differences in Pol II (a surrogate for transcription level) are eliminated.

3.4.6 Splicing Signal Hidden Markov Model

To define the hidden Markov model (HMM) for splicing signals, we start with a 5 splice site (SS5) position weight matrix (PWM) estimated from GENCODE 16 annotations of the first exon for protein-coding genes (Figure B.29b). In addition, a PWM for poly(A) sites (PAS) was estimated from the sequences reported in Beaudoing et al. [12]. Finally, a background model was estimated from the full DNA sequences, assuming independence of sites.

Our HMM combines these motif models in such a way that we can make inferences about the relative positioning of SS5 and PAS sequence motifs. In particular, the HMM permits branching from an initial background state into five alternative paths. Two of these paths visit the SS5 site before an optional PAS; two others visit the poly-A site before an optional U1 site; and a final path includes none of the two motif signals. The HMM is structured such that the transition from the initial background state is taken once and only once (Figure B.29c).

We applied this HMM to sequences spanning the first 1.5 kb of TSSs in each class (stable and unstable). To estimate the relative likelihood of each path, we computed maximum likelihood estimates of the transition probabilities into each of the five alternative paths using the Baum-Welch algorithm [34]. Because the number of free parameters is the same for all paths, no model complexity penalty is needed for this comparison.

Additionally, the probability of each alternative path for each sequence can be estimated by setting the uniform transition probabilities out of the initial background state and then computing the respective the posterior probabilities. This enables the use of the HMM model as a sequence classifier (by thresholding the sum of the posterior over the sequence) and it is used bellow as the input for the stability regression.

3.4.7 Stability Regression

The relative contribution of individual TFs and the splicing signal HMM towards predicting the TSS class (stable or unstable) was assessed by logistic regression. TF signals correspond to sums of ChIP-seq signal in the predicted TSS region. Relative importance of regression weights was computed according to Johnson [70]. Because TF binding patterns are often strongly correlated with transcription level, we applied the logistic regression to subsamples of stable and unstable TSSs with matching Pol II signal distributions (as described above).

CHAPTER 4

UNSUPERVISED TRANSCRIPTION UNIT IDENTIFICATION

4.1 Introduction

The pursuit of a deeper understanding of transcription regulation has led to the recent development of experimental methods that map genome-wide transcriptionally engaged RNA polymerases (RNAP) (GRO-seq [27], NET-seq [23] and PRO-seq [81]). These assays take advantage of short-read sequencing and specialized chemistry to precisely map the edge of nascent RNAs, therefore producing density profiles of RNAP across the genome. These profiles reflect the underlying dynamics of transcription, including effects such as promoter-proximal pausing and waves of transcription in post-induction time-series [55]. Fundamentally, these assays enable the comprehensive detection and analysis of both traditional transcription units (protein-coding genes), and less well understood ones, such as long intergenic RNA (lincRNAs). Importantly, nascent RNA based assays also capture rapidly degraded transcripts (e.g., divergent transcription, enhancer RNAs (eRNAs)), a subclass not easily accessible to assays based on spliced/accumulating RNA, such as RNA-seq, without requiring changing in-vivo conditions (e.g. exosome depletion [109]).

The effective identification of transcription units (TUs) from datasets produced by GRO-seq (and related assays) is not adequately covered by existing tools designed for RNA-seq, due to the intrinsic differences in the datasets produced by these two types of assays. RNA-seq, by virtue of being based on spliced accumulating mRNAs, produces a reasonably uniform sequenced read distribution (mRNAs are randomly fragmented prior to sequencing) over a rel-

atively small portion of the genome[1]. These features, together with paired-end sequencing, enable methods such as Cufflinks [116] to produce and join contigs (regions obtained by merging partially overlapping reads) into full transcripts. In contrast, GRO-seq reflects engaged RNAP positions, estimated to be on average one per cell per transcript [66], that span the entire TU, and furthermore reflect dynamic aspects of transcription (e.g. pausing) and possibly synchronization across the sampled cell population. As such, GRO-seq produces a sparse, non-uniform, read density signal requiring specialized methods for TU identification.

TU identification is essentially a genome segmentation problem, dividing regions into transcribed and non-transcribed. Naturally, one of the initial methods for GRO-seq data made use of a two state hidden Markov model (HMM) to perform this segmentation [55]. Although some of the model parameters were set via Expectation Maximization (EM), good model performance required supervised parameter tuning based on existing RefSeq gene annotations. A subsequent method attempted to improve TU identification through a contig based algorithm, with contig joining performed as a function of density and distance [2]. Parameters were tuned based on an external transcription initiation dataset, but even then, performance only surpassed the previous HMM method when RefSeq annotations were used to control contig merging. In either case, good performance was dependent on existing annotations, which limits the utility of these methods.

We improve on previous work by developing an HMM based model that does not require annotations for parameter estimation, expanding the useful-

[1]RNA-seq only covers exons, which account for about 3% of the main protein-coding genes [119], and does not cover transcription beyond the poly-adenylation site, nor unstable transcripts.

ness of this class of models. Furthermore, we present an approach to integrate transcription initiation signal, when available, directly in the HMM model. These results are evaluated on a collection of GRO-seq and PRO-seq datasets, on both human and *Drosophila* cell lines, with a wide range of read densities. To perform this evaluation, we combined GENCODE annotations with additional data sources to obtain cell specific reference sets. Finally, TU predictions on the K562 cell line are combined with polyA-seq to exemplify the application of TU predictions to the characterization of the later stages of transcription and processing.

4.2 Results

4.2.1 Hidden Markov Model

In order to take advantage of nascent RNA based assays (such as GRO-seq and PRO-seq) to identify transcription units, we applied a two-state hidden Markov Model (HMM) to a transformation of the read count data. This data transformation splits the raw, per position (or step), read counts into a sequence of non-zero read counts and distances to the next non-zero position (or step) (see Figure 4.1a,b). HMMs using the data transformation have consistent high F1 values (the F1 measure summarizes the balance between precision and recall) across the wide range of library densities (2 to 143 reads per thousand mappable bases), with largely better values than equivalent non-transformed HMMs (see Figure 4.1c, Table 4.1; see Methods for details on evaluation metrics). Furthermore, the fraction of errors per matched reference TU (errors consist of either

Table 4.1: Nascent RNA dataset information. (*) Maximum read length and value used to compute mappability. In some cases smaller reads were also mapped. (#) Separate per base sequencing, read counts after normalization. Rounded to nearest integer.

Dataset ID	Dataset	Type	Read Length(*)	Reads per thousand mappable bases
A	hg19/hela	GRO-seq	30	2.17
B	hg19/k562	PRO-seq	50	3.27
C	hg19/k562	GRO-seq	30	6.78
D	hg19/imr90	GRO-seq	30	12.60
E	hg19/mcf7	GRO-seq	44	32.48
F	hg19/gm12878	GRO-seq	30	39.60
G	hg19/k562	PRO-seq	100	127.42
H	dm3/s2	GRO-seq	26	116.32
I	dm3/s2	PRO-seq (#)	26	143.78

over fragmentation or incorrect transcript merging) is also consistently lower than that obtained by using equivalent non-transformed HMMs, and largely insensitive to library densities (see Figure 4.1d). We experimented with a variety of emission distributions (data not shown) with the best performances obtained combining the use of the Geometric distribution for distances with some combination of Poisson, Discretized Gamma (as used in Hah et al. [55]) and Negative Binomial. Noticeably, the data transformation reduces the effect of read count distribution choices (see Figure 4.1c,d), showing a small advantage for the combination of Poisson and Negative Binomial distributions, for background and transcribed states respectively, which we will use henceforth.

TU detection performance can be improved by incorporating a signal for TSS locations. We opted to integrate this additional information through the transition probability function from background to transcribed states ($P(Z_i = B|Z_{i-1} = T)$), turning the model into a non-homogenous HMM conditional on the TSS signal. When defining this probability distribution, it is necessary to balance the signal to form a new TSS with the potential for over-fragmentation

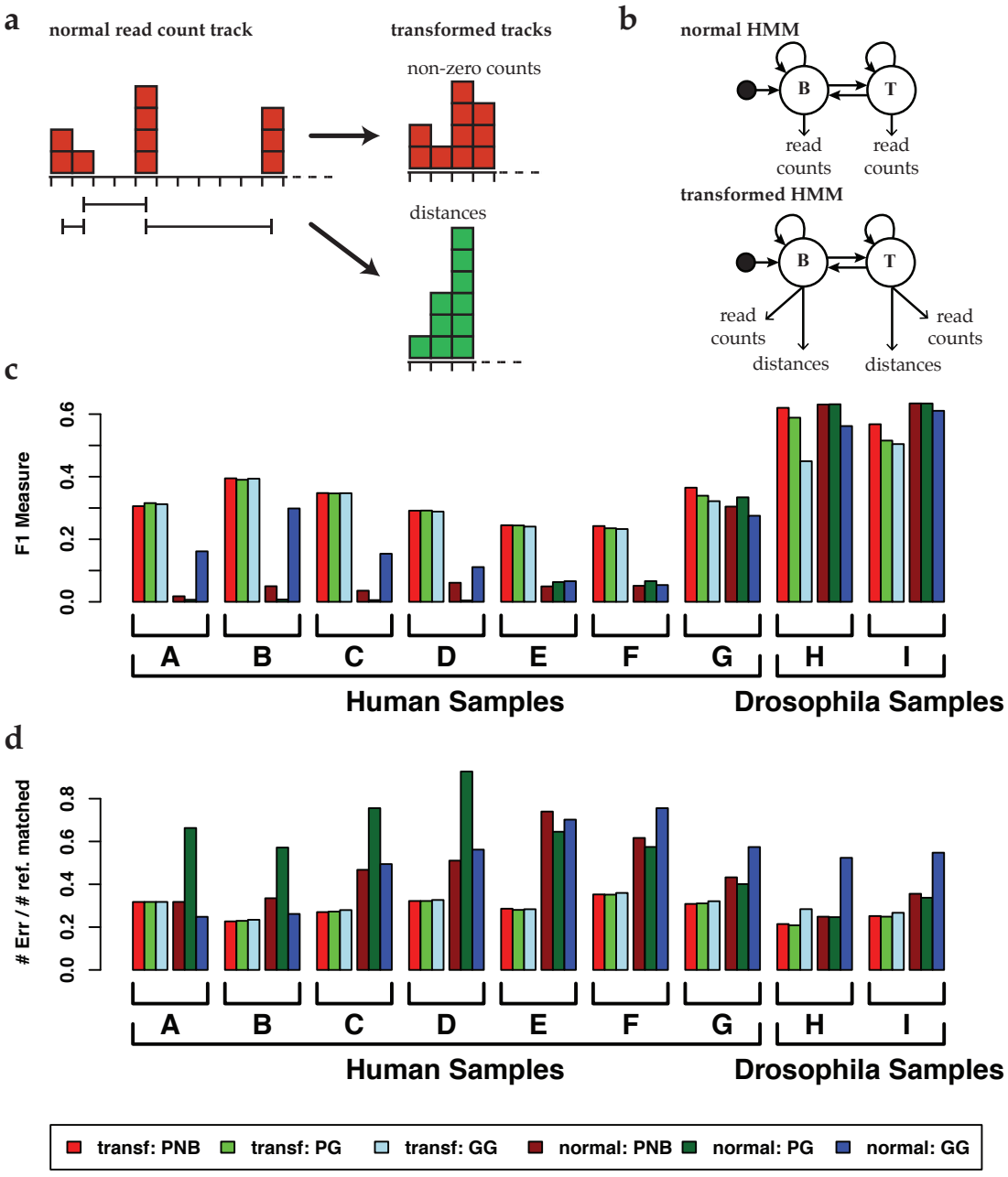

Figure 4.1: TU HMM data transformation. (a) Read count track is transformed into two tracks: (1) non-zero read counts and (2) distance between non-zero read count positions. (b) Two-state HMM model variants (B: background, T: transcribed): normal HMM (top) states emits read counts and transformed HMM (bottom) states emit both non-zero read counts and distances to the next non-zero read count position. Performance metrics on datasets from Table 4.1, (c) F1-measure (harmonic mean of precision and recall; higher is better) and (d) fraction of errors per matched TU prediction (transcript merge and fragmentation errors; lower is better), for each combination of normal or transformed HMM and choice of read count emission distributions (B,T): PNB (Poisson,Negative Binomial); PG (Poisson,Discrete Gamma); GG (Discrete Gamma,Discrete Gamma).

that results from overlap between transcripts (e.g. in-gene enhancers). Furthermore, for a general use method, it is desirable to have a simple way to exchange alternative TSS data sources. We tried two alternative formulations, either a fixed parameter γ or a function of the TSS region score (see Methods). In Figure 4.2a,c we show the results with two TSS region sets, GRO-cap predicted TSS regions (results from Chapter 3) and TSS regions obtained from a computational method (dREG [31]), in the three datasets where both sources are available. Overall, adding the TSS information improves performance, both in terms of precision/recall (summarized in the F1 measure) and in terms of the number of errors. Effects are more strongly felt with lower density samples and are only marginally affected by the choice of the γ parameter (a low 0.1 value seems like a safe default choice). As expected the GRO-cap dataset performs better (and is less affected by the γ parameter), being derived from a direct experimental measure of TSS positions, but the computational method is very competitive. Incorporating the score (from the dREG source) by itself has negative consequences (see Figure 4.2d).

Integration of TSS signals, as described, provides enough constraints to parameter optimization to permit the useful expansion of model complexity while keeping with annotation free parameter estimation. As can be seen in the example of Figure 4.3, some transcripts have a prolonged post-polyA transcription tail. When these transcripts are packed closely to another transcript on the same strand this leads to incorrect merging of adjacent transcripts. We aim to counter-act this issue by the addition of an extra decay state (see Figure 4.2b and Methods), i.e, an extra state, reachable from the main "transcribed" state, that aims to identify the low level transcription that extends beyond the end of the main transcribed region. Without the extra TSS signal, this extension leads to

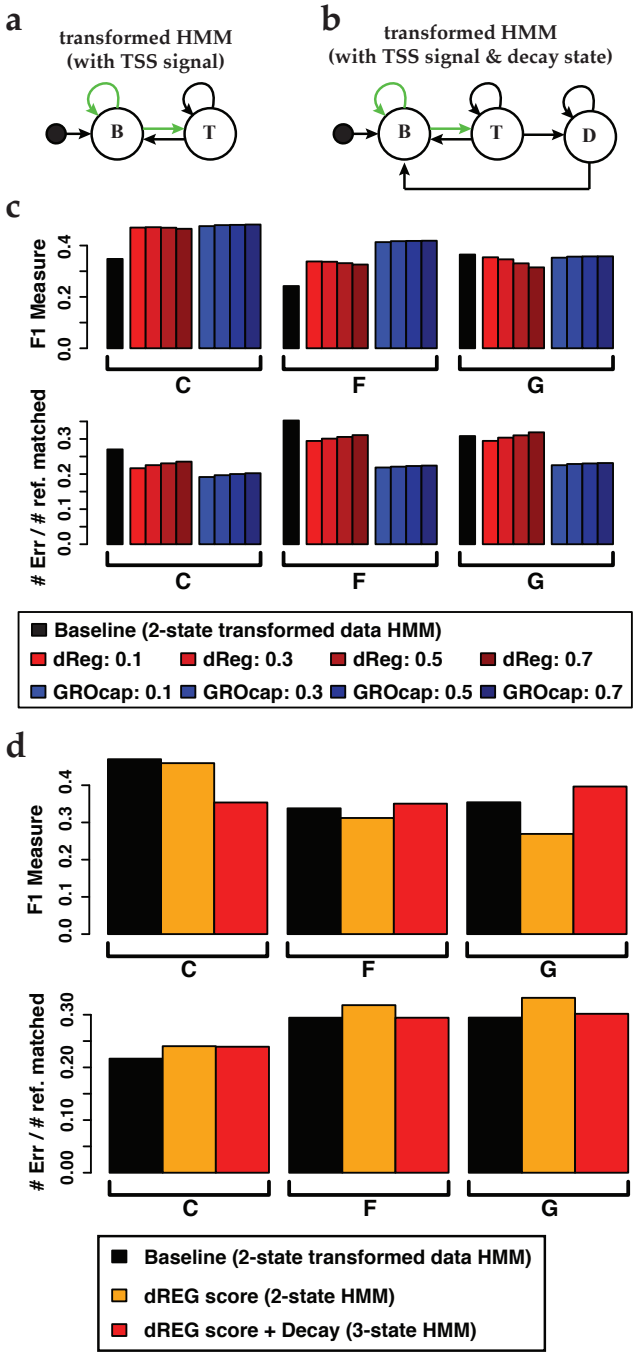

Figure 4.2: Extending the TU HMM. Two variations on the base HMM: (a) transitions from background state B are defined by external signal; (b) TSS signal and additional decay state. (c) Two different TSS signal source (GRO-cap, in blue, and dREG, in red, baseline in black) are tested with different values of the γ parameter in the human datasets where both are available. Performance metrics, F1-measure (harmonic mean of precision and recall; higher is better) and fraction of errors per matched TU prediction (transcript merge and fragmentation errors; lower is better), for each combination of source and γ value. (d) Comparison of HMM extensions using dREG score as the probability (see Methods) using the same performance metrics: F1-measure and fraction of errors per matched TU.

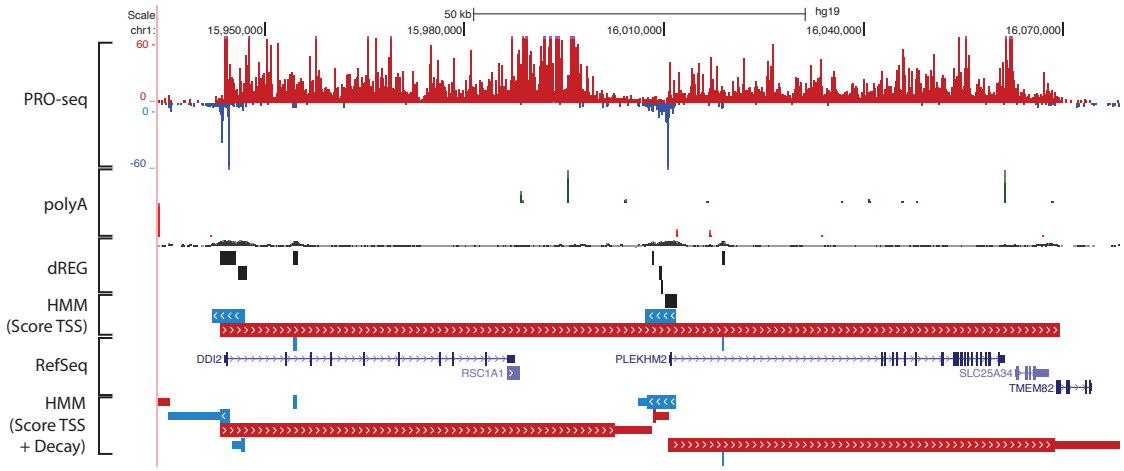

Figure 4.3: TU incorrect merge example (UCSC Genome Browser screenshot). Tracks show data for the G dataset (K562) (PRO-seq, polyA-seq, dREG) together with RefSeq annotations and two sets of TU HMM predictions: 2-state with dREG score as TSS signal and 3-state with dREG score as TSS signal and extra decay state. Example illustrates neighboring gene signal bridging due to post-polyA extended decay.

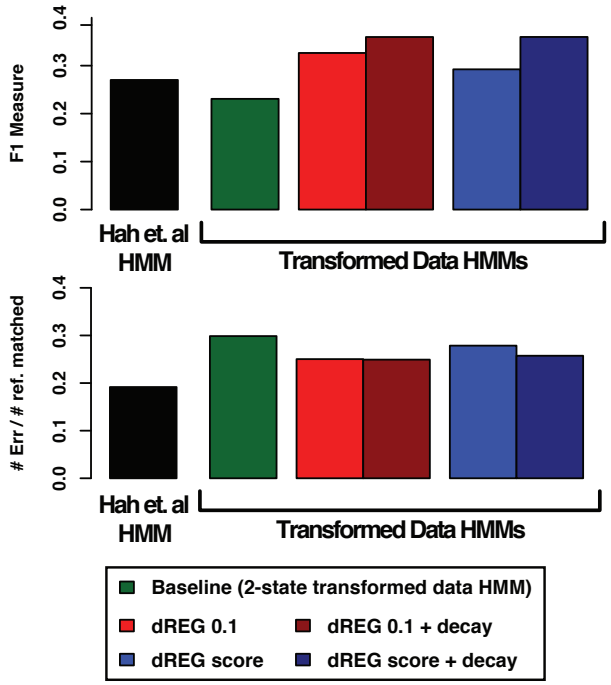

Figure 4.4: TU HMM variants comparison. Unsupervised HMMs using the data transformation achieve comparable performance with similar error rates to the HMM from Hah et. al [55]. Addition of dREG as TSS signal and the extra decay state further improve performance and reduce error rate.

degenerate behavior (see Figure C.1a). Even with the simplified single parameter TSS signal version, the results are not satisfactory (see Figure C.1b). However, in conjunction with the dREG score it improves performance on moderate or high GRO-seq/PRO-seq library sizes (see Figure 4.2d, Figure 4.3), reversing the negative impact that the dREG score had by itself.

Overall, the data transformation results in comparable performance to the HMM by Hah et. al [55] which made use of existing annotations to tune the model parameters (see Figure 4.4). The inclusion of the GRO-seq derived TSS signal, from dREG, with the extra decay state, improves the performance, as measured by the F1-measure, beyond the annotation driven HMM, while reducing the gap in terms of fraction of errors per matched transcript. In the following, we explore some aspects of transcription unit profiles using the predictions obtained on *G* dataset (high density K562 PRO-seq) with the 3-state HMM incorporating the dREG score based TSS signal.

4.2.2 Transcription beyond the poly-Adenylation site

Contrary to RNA-seq, nascent RNA assays such as PRO-seq capture transcription beyond the site of cleavage and poly-adenylation (polyA) of the primary transcript. We took advantage of the transcription unit predictions from our HMM to model the variation of the post-polyA extension across genes. Using published polyA-seq data on K562 cells [88], together with deep sequenced PRO-seq, we selected TU with at least one polyA cluster (picking the cluster with the highest read count) and divided them into three regions *body*, *post-polyA pause* and *decay* (see Methods). Profiles aligned at the polyA site show a

strong PRO-seq signal accumulation over a wide region (see Figure 4.5a). This was previously reported as pause induced by the poly-adenylation process [27], but further study found this to be inconsistent with the profile expected from a single pause source [52]. The *post-polyA pause* extends on average 10kb (median 5kb), while the estimated decay is on average 23kb long (median 10kb) (see Figure 4.5b). This wide range suggests that other factors beyond simple polyA site induced pausing may be at play. Examination of signal intensity, at the three stages (before polyA, during the pause region and afterwards) shows strong correlation of signal intensities, with linear relationship between pre-polyA signal and signal both at the pause region and afterwards (see Figure 4.5c). We further classified TUs into ”Paused” and ”Not Paused” by computing a confidence interval on the pause to body ratio (assuming counts follow a Poisson distribution [35]). This results in 62% (out of 9077 TUs) that are significantly paused (95% confidence interval of the ratio above one).

Curiously, the ”Paused” class has a significantly higher mean gene body level (one sided Wilcox test p-value $<$2e-16) than the ”Not Paused” class, raising the possibility that a non-linear effect related to cross-strand collisions, or a feedback effect with promoter-looping, may be at play. Cross-strand collisions between engaged polymerases can result in excessive pausing, although we found no evidence for this to be the main mechanism controlling *post-polyA pause* level (a small cross-strand effect can be glimpsed in Figure C.3, where the strong *post-polyA pause* on the negative-strand produces a small local increase on the plus strand transcription level).

CTCF is one of the TFs responsible for establishing distal chromatin links. This has led to a variety of regulatory roles being ascribed to it, depending on

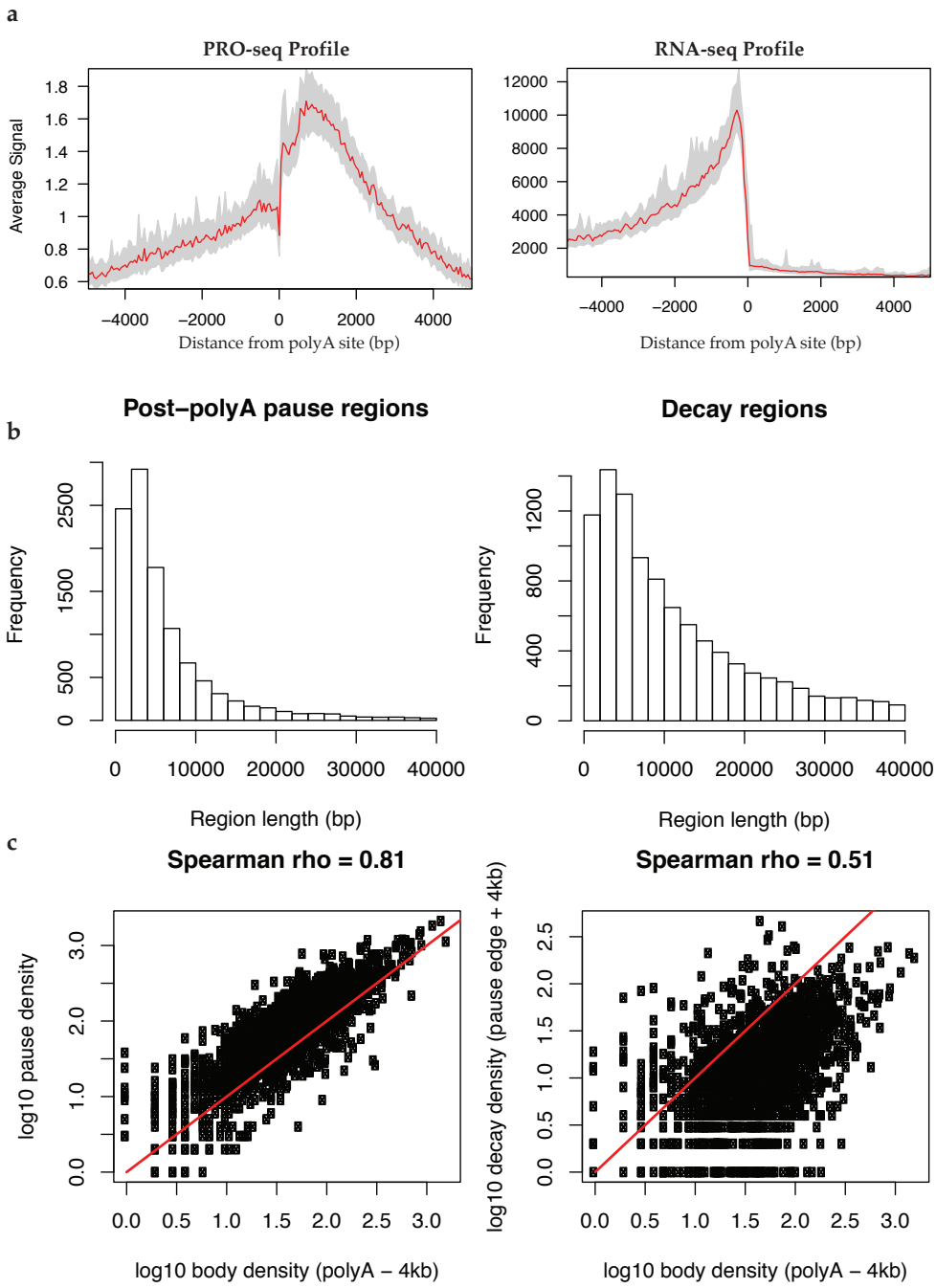

Figure 4.5: Features of TU decomposition. TU were split into three regions, *body*, *post-polyA pause* and *decay*. (a) PRO-seq and RNA-seq profiles aligned to the polyA site. (b) Length distributions (truncated at 40kbp) for *post-polyA pause* region and *decay* region. (c) Scatter plot of log10-density of *post-polyA pause* versus *body* (left) and *decay* versus *body* (right).

the regions that are linked [98]. Therefore, it follows that it may play a role when present at the edge of a post-polyA pause region. However, although CTCF is prolific across the genome, only 4% of the selected TUs contain a well placed CTCF element at the edge of the post-polyA pause region (198 in the "Paused" class, or 3.5% and 99 in the "Not Paused" class, or 2.8%). Furthermore, we found no evidence to support that this subclass behaves differently across an induction time series (e.g. Celastrol; data not shown).

4.3 Discussion

We presented an unsupervised approach to predict transcription units (TU) from nascent RNA (GRO-seq/PRO-seq), as well as an approach to perform evaluation of the model's performance. Key to this advancement is the data transformation, which enabled stable parameter inference across a wide range of library densities. We further explored extensions to the simple two-state model, incorporating TSS signals from experimental assays (GRO-cap) and from synthetic methods (dREG). Finally, we demonstrate the usefulness of these predictions by analyzing post-polyA transcription extension and showing that the paused region extends considerably, 10kb on average; in many cases, a further low level decay extends the reach of the TU even further.

4.3.1 Data transformation

The data transformation used as a basis for our TU HMM reduces the number of transitions over minimally informative positions (zero-read count positions

which are abundant due to the sparse nature of GRO-seq and similar assays), reducing the excessive transcript fragmentation observed in previous work without requiring a supervised parameter tuning [55]. Although in the absence of noise, both measures are strictly related (under simple Poisson assumptions of read count distribution), actual data incurs both an excess of zero-read-count positions, due to both the intrinsic nature of the assay and sequence mappability constraints, and occasional read spikes due to uneven read amplification. The added freedom provided by the data transformation, provides increased tolerance for these issues and in particular unmappable gaps, as these now translate into sporadic excessively long distances and thus have less impact on path inference. Furthermore, the data transformation provides a minimal degree of non-locality to the model, which may contribute to its improved performance.

An interesting perspective on the data transformation is the parallel that can be drawn between the resulting HMM and a continuous-time HMM, in particular, it is very similar to a discretized version of the continuous-time HMM framework. The distance (or time) until the next step is modeled by geometric distributions, which can be seen as discretized versions of the exponential distributions used in the continuous-time HMM framework. Thus, for the homogenous version of the models (no TSS information), we can rewrite the HMM in a way that parallels the continuous-time HMM likelihood. This observation provides a possible approach to bridge these models with the underlying physical process, as contrary to RNA-seq assays, nascent RNA based assays like GRO-seq and PRO-seq, map the positions of engaged polymerases across the genome. The relatively small number of polymerases per gene [66], together with ideas such as transcriptional bursts [112], may offer a path for further exploration of these observations.

4.3.2 TU in the genomic context

The brief analysis of post-polyA TU extension hinted at several issues that affect not only the transcription landscape but may need to be better understood to improve the model. Issues such as chromatin linking (e.g. via CTCF), which may result in complex feedback between distal areas of the genome or simply act as a barrier, will necessarily affect the observed transcription profiles. It is noticeable that TUs with strong post-polyA pausing also tend to have a higher activity level in the gene body, this may be the result of promoter looping as discussed in Grosso et. al [52]. Furthermore, cross-strand collisions between engaged polymerases leave their footprints on TU profiles, in the form of local increases in observed GRO-seq/PRO-seq levels. This can potentially result in incorrect segmentation from simpler models that expect cleaner profiles. These challenges should be kept in mind as the field advances towards a higher resolution modeling of the biology underlying genome-wide datasets.

This work shows that both different perspectives on the data and the integration of external constraints (like the TSS signal), open the opportunity to increase model complexity without sacrificing performance. It is clear that there is room for the integration of additional, specialized, signal detectors (e.g. for post poly-A decay), as well as for expanding the model to incorporate inter-TU interactions, either cross-strand or due to isoform overlap (e.g. via some form of TSS driven flow diffusion model). Nevertheless, the current state of TU models is at the point were they are useful components in downstream data analysis projects.

4.4 Methods

4.4.1 Transcription unit evaluation

Active transcription units vary from cell type to cell type and across experimental conditions. Genes, for example, are usually annotated as groups of alternative transcripts with different TSSs, exons and polyA sites, with the active subgroup varying across conditions. It is therefore necessary to combine condition specific assays with known annotations to produce a reliable validation set.

Moreover, different types of transcription units require different validation strategies and datasets. We choose to split transcription unit types in two sets: long spliced TUs and short transient ones. In the long TU set we find protein coding genes (and their degenerate forms, such as pseudogenes) and lincRNAs. This set is readily visible in RNA-seq and CAGE assays and is widely annotated up to the poly-adenylation (polyA) site. The second set is formed mainly by short enhancer driven transcription and short divergent transcripts at promoters of the first set. The transcripts of the short TU set are subject to rapid degradation and do not show on RNA-seq or CAGE assays. Moreover, not only is their annotation much less precise but they are also more variable across conditions. These two sets require different evaluation strategies.

To further complicate matters, transcription units of the short TU set, namely transient enhancer transcripts, can be contained within transcription units of the long TU set.

Base pair precision in TU prediction validation is impossible, as should be

clear by now. Instead, each reference set imposes restrictions on the start and end of a TU with respect to a matched reference TU. Furthermore, restrictions are also imposed on how different TUs can overlap with each other.

Overall, the evaluation strategy can be split into three parts: 1) reference set definition; 2) matching algorithm; 3) evaluation metrics. Which we approach in turn in the following.

Long spliced reference set

This set of transcription units has good annotation coverage up to the polyA sites and good assay coverage in ENCODE/modENCODE cell lines. It is characterized in annotations in groups of alternative forms, sharing common DNA. So, we define a *correct* TU prediction as one that satisfies the following conditions:

- Only overlaps with a single reference group.
- Starts within T_S bases of the reference TSS.
- Extends at least until the annotated polyA site.

We define this reference set by combining the information from three sources (data and values for GM12878, see Table C.1 for full reference sets)[2]: GENCODE 16 annotations, ENCODE Long polyA+ RNA-seq and ENCODE CAGE.

Starting with the GENCODE transcript annotations, we selected those with at least two exons (spliced transcripts) and a minimum amount of RNA-seq read density in each constituent exon. Two possible ways to set the threshold are to

[2]These sources are for Human assays, but similar sets could be used for other species.

maximize the number of transcripts per gene group or to maximize the average number of exons per transcript (see Figure 4.6). We chose to set it to maximize included transcripts (log(RNA-seq density) $\geq$ 0), as these would be subject to further filtering steps and it leads to a larger number of gene groups (14493 vs 11979, in the GM12878 dataset).

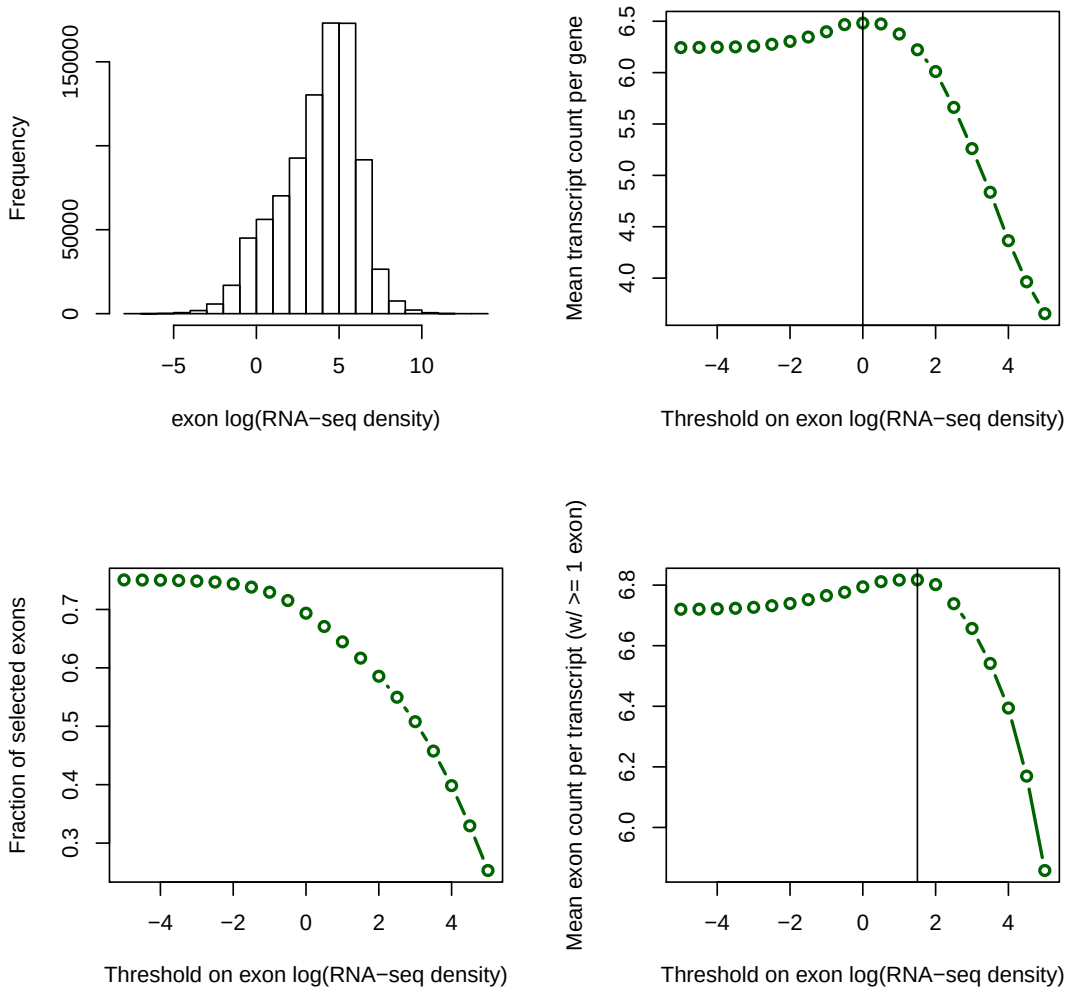

Figure 4.6: Annotation selection via exon RNA-seq density threshold (GM12878 cell line).

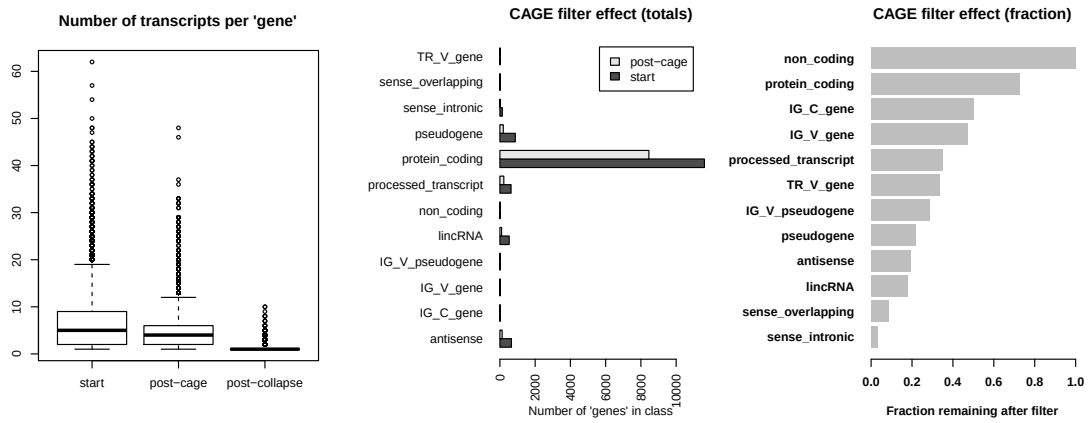

Figure 4.7: Effects of the various filters on transcript annotations (GM12878 cell line).

The resulting annotations were further filtered by requiring a minimum of 50 CAGE reads within 250 bp of the annotated TSS. This brings the number of gene groups down to 9107 (with 44585 transcripts; in the GM12878 dataset). Finally, within each group, transcripts that started within 250 bp of each other were collapsed, resulting in 11565 transcripts (in the GM12878 dataset). The reduction of transcripts per group and the representation of each GENCODE transcript type can be seen in Figure 4.7. This filtering step resulted in sharper signal profiles for relevant assays, including both promoter and gene body marks (see Figure 4.8).

Short transient reference set

The construction of a reliable short transient TUs set is a much harder problem. By their very nature, transient transcripts do not show up in RNA-seq or CAGE assays. They are generated at active enhancers and in the divergent sense of gene promoters, though not all such transcripts are transient (eg. lincRNAs). Adequate identification of transient transcript boundaries is further complicated by the tendency of enhancers to occur in closely packed regions

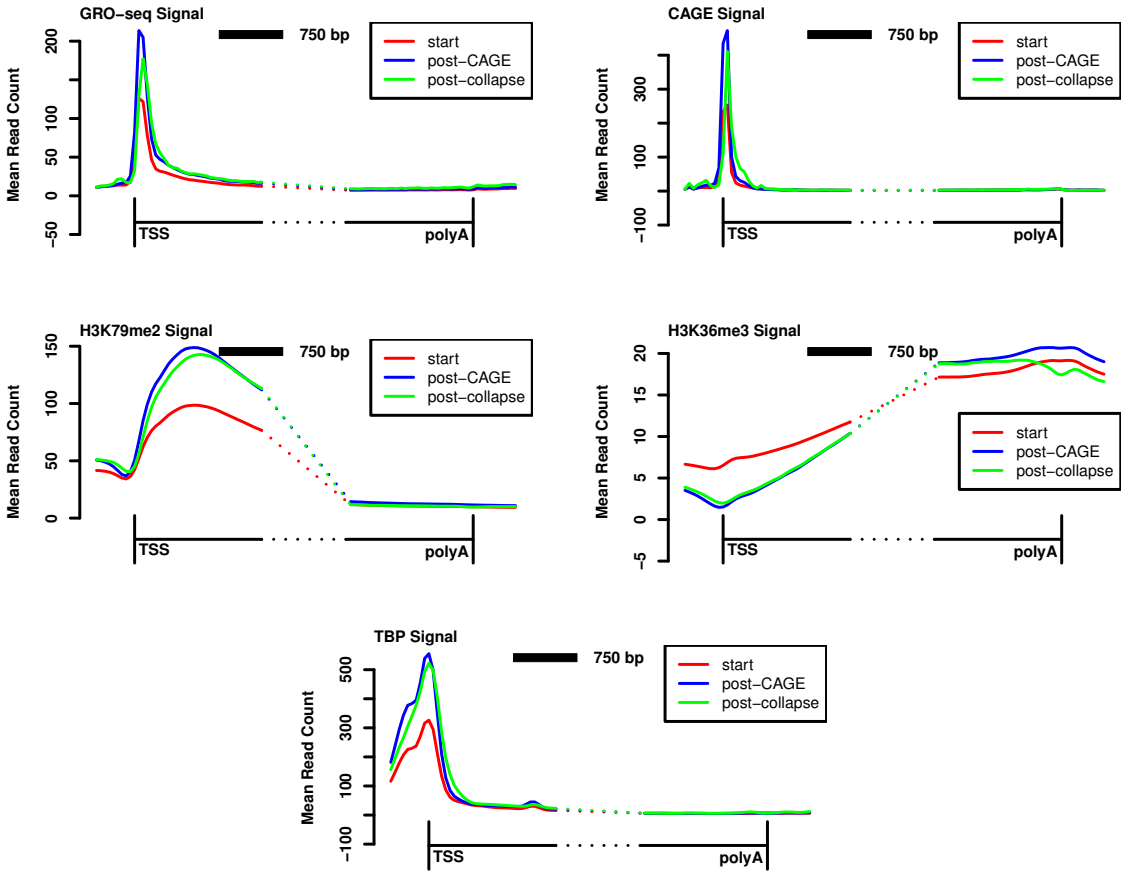

Figure 4.8: Profiles of several characteristic gene signals at successive filtering stages. Start here represents the transcripts selected via RNA-seq thresholding (GM12878 cell line).

and to overlap gene transcripts.

Short transient transcripts are, nonetheless, initiated in a similar fashion to regular transcripts (see Chapter 3) and so their presence is revealed in chromatin assays such as DNAse HS, histone modification ChIP-seq. H3k27ac is a known mark of transcription initiation and known selector of active enhancers [29][97]), therefore, we selected as the reference set, the DNAse HS peaks that overlapped with the H3k27ac peaks (see Table C.1).

As described, not much information exists on the precise definition of short

transient TU boundaries; so, at a basic level, predicted TUs can be called "correct" when overlapping with these regions (exceptions, when the reference short TU is overlapping a reference long TU region, are noted in the next sub section). If a more strict test is required, we can require that they start within such a reference region.

Reference matching algorithm

To evaluate a prediction set, it is necessary to mach elements in that set with those in the reference sets. This will label the elements in both sets with one of three labels: *correct*, *incorrect* or *unknown*. The totals for each set are used to define the evaluation metrics described in the next section.

The description of the reference set, in particular the subset of *Long spliced TUs*, allows for groups of overlapping alternative TUs (genes, for example, have multiple alternative transcripts). Since not all TU prediction models allow for overlapping TUs, we define two match modes: a *strict* mode, where all alternative TUs are considered independently and a *relaxed* mode, where each group is considered as a single unit. Furthermore, in relaxed mode, predictions are allowed to break at inner TSS boundaries within a group and are allowed to merge alternative TUs within a group.

The match process starts by creating, for each element in either set, a list of the elements in the opposing set that overlap (on the same strand). Note that if we are conducting the match process in the *relaxed* mode, then the reference set is enriched with additional TUs per group that correspond to the relaxed conditions (breaks at TSSs and merged alternative TUs). Given this information,

the process is composed of two phases. In the first phase, a sequence of labeling steps is applied to the prediction set, where the input of each step is the set that was not labeled in the previous steps. In the second phase, labels of the prediction set are used to define the labeling of the reference set. The prediction set labeling steps are:

1. Elements that have empty overlap lists are labeled as *unknown*.
2. Elements that overlap more than one TU group in the *Long spliced TU* reference subset are labeled as *incorrect* due to over-extension.
3. Elements that match one or more TUs in the *Short Transient TU* reference subset (and only in this subset) are labeled as *correct*[3].
4. Elements that match one and only one reference TU are marked as *correct* or *incorrect* according to the rules defined in the previous sections.

At this point, we are left with predictions that overlap only a single isolated group of *Long spliced TUs* and zero or more *Short transient TUs*.

5. Given a particular prediction TU, check if a correct match can be found against each reference TU in the group of *Long spliced TUs*, in order of size (check from longest to shortest)[4]. If a match is found, the prediction TU is labeled as *correct*. The only exception is matching against TU fragment added when augmenting the reference set in *relaxed mode*. In this case, the prediction TU is only labeled as *correct* if at least one other reference TU starting at the TSS that lead to the break is also labeled *correct*[5].

[3]Should we desire to be more strict, require a prediction TU to start within a particular reference TU and only count it towards that TU.

[4]Preference is given to *Long spliced TUs* in the reference set as those are more reliable.

[5]This is done to prevent an incomplete prediction to be mistakenly labeled as correct.

6. Given a particular prediction TU, T, if it is covered by another prediction TU[6] and it overlaps a *Short Transient TU* for which there is an opposing strand prediction TU that starts at that *Short Transient TU*, then mark T as *correct*[7]

All remaining unlabeled prediction TUs are labeled as *incorrect*.

Labeling the reference set is done optimistically, that is, a reference set element is assigned the best label of its overlapping (matched, ie, used to call correctness) prediction TUs. Elements with no overlapping prediction TUs are labeled *unknown*. In *relaxed* mode, groups of TUs are combined under a single label (*correct* if any of the component TUs is *correct*).

Evaluation metrics

Given the result of the matching algorithm, we now define the evaluation metrics based on the total counts of *correct*, *incorrect* and *unknown* elements: CR, IR, UR and CP, IP, UP for the reference and prediction sets, respectively.

The first metric is the *F1-measure*, used in *information retrieval* and defined as the harmonic mean[8] of *precision* and *recall*. To that end, we adapt the traditional definitions of precision and recall to take into account *unknown* labelings:

$$\begin{aligned} \text{precision} &= \frac{CP}{CP + IP} \\ \text{recall} &= \frac{CR}{CR + IR + UR} \end{aligned}$$

[6] Avoid labeling parts of a fragmented gene prediction as correct just because it overlaps an internal enhancer.

[7] Collision between opposing strands results in increased residence time, which intensifies the signal of either transcript (the effect size is the subject of further study). As such, we can expect the opposing strand to be detectable if the sense strand was a true match.

[8] Reciprocal of the arithmetic mean of the reciprocals.

The lack of symmetry with respect to the treatment of the *unknown* label is a reflection of the incomplete nature of our reference set. We can be reasonably confident that the TUs of the reference set are indeed present in that experimental condition, thus they should be counted when computing *recall*. However, we do not have a exhaustive reference set and thus cannot make a judgment call over predicted TU labeled as *unknown*. So, we have chosen to be optimistic in defining *precision*.

The second metric is the fraction of errors per matched reference. Here, errors are defined to include merging of adjacent transcription groups (determined when applying step 2 of the matching algorithm) and over fragmentation (determined when labeling references in the second phase of the matching algorithm). This metric aims to provide information on the quality of the overall prediction set, based on the subset that matched with known references.

4.4.2 HMM Parameter Estimation

Maximum likelihood estimates of the free parameters are obtained via Expectation Maximization applied to the various datasets (tracks using 50 bp steps). In practice, the estimation is performed per chromosome, enabling trivial parallelization of the process. Moreover, parameter estimates from a small chromosome (chr22 in humans, chr4 in Drosophila) are used as the initial guess for the remaining chromosomes, speeding up the process.

4.4.3 Encoding TSS Information

The TSS signal is incorporated in the HMM model as the probability of a transition from the background state B to the transcribed state T, $P(Z_i = T|Z_{i-1} = B)$. We test two alternative approaches. The first is a simplified version, where the only information provided are TSS regions (in a BED file) and a user defined value, γ. Locations inside these regions are assigned γ as their transition probability and locations outside have zero probability of taking that transition. That is:

$$P_{BED}(Z_i = B|Z_{i-1} = T) = \begin{cases} \gamma & \text{if position } i \text{ is in a TSS region} \\ 0 & \text{otherwise.} \end{cases}$$

where γ is a user supplied parameter. The other alternative we considered is, when given a score associated with each TSS region, transform that score into a probability:

$$P_{score}(Z_i = B|Z_{i-1} = T) = \begin{cases} f(score_j) & \text{if position } i \text{ is in TSS region } j \\ 0 & \text{otherwise.} \end{cases}$$

In the dREG case, scores are the output of an SVN where the TSS sites have the value one as the label and non-TSS sites have the value zero as the label. As such, we take the heuristic approach of using the score directly by clamping it: $f_{dREG}(score_j) = min(1, score_j)$.

4.4.4 Post PolyA Decay Extension

Post polyA transcription, after the short tail of similar or higher transcription level to the main gene body, typically has a very low uniform expression level. Therefore, we model it with a Poisson distribution, similar to what is done with

the background state. As this can potentially lead to loss of sensitivity (low level transcripts absorbed into the decay state), we tested HMMs with multiple transcript paths bound together to share parameters up to a scale factor (see Figure 4.9). In the case of the Poisson distribution the scale factor is trivial to incorporate, just decompose the Poisson rate into a common factor and a fixed scale value. For the Negative Binomial emission, we apply the scale factor to the "dispersion parameter" r. We use the following form for the Negative Binomial:

$$P(X = x|p, r) = \frac{\Gamma(x + r)}{\Gamma(x + 1)\Gamma(r)} p^x (1 - p)^r.$$

Replacing r with αr, where α is our scale factor, has the effect of also linearly scaling the mean ($\mu = \frac{pr}{1-p}$), which works as intended and is consistent with the behavior of the Poisson distribution.

Transitions from the background state B to any of the T_j states are equiprobable, so for example, if that probability is defined by γ as described above, then $P(Z_i = T_j|Z_{i-1} = B) = \gamma/K$ and $P(Z_i = B|Z_{i-1} = B) = 1 - \gamma$.

4.4.5 Refined TU Regions

To prepare the TUs for further analysis, we started with the TUs produced by the 3-state HMM with dREG score based TSS information. These were then filtered for TUs with at least 5 kbp, excluding the decay region, since we want to analyze post-polyA pausing and decay and that region is expected to be at least 4 kb wide. These were then filtered to select those that had at least one polyA-seq cluster, resulting in 11196 TUs. Thus, TUs were split into main body (start up to polyA cluster), pause region (polyA cluster up to main body end) and decay (HMM predicted decay end).

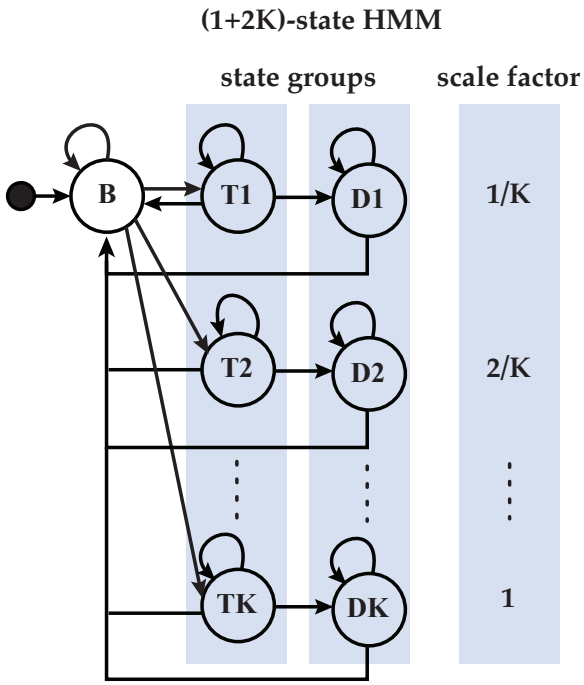

Figure 4.9: Extended TU HMM with multiple paths. All states in each group (T_j and D_j) share the same underlying emission parameters, but have their own scale factor j/K. This constrains the HMM to keep the same relation between transcribed and decay states to avoid overfit.

Although the HMM's decay state helps approximate the edge of the main part of the TU to the start of the decay, it is imperfect. To ensure the best results in downstream analysis, we refined it by scanning from the polyA site to the HMM estimated start of the decay edge + 1kb (some TUs don't have any decay estimated). Assuming Poisson distributions, we picked the position that maximized the log-likelihood of the PRO-seq read count probability in the two-part region (each part has its own lambda, set to the region mean read count). See Figure C.2 for effect of refinement on PRO-seq and RNA-seq profiles, and Figure C.3 for a browser shot example.

APPENDIX A

SUPPLEMENTAL MATERIAL FOR CHAPTER 2

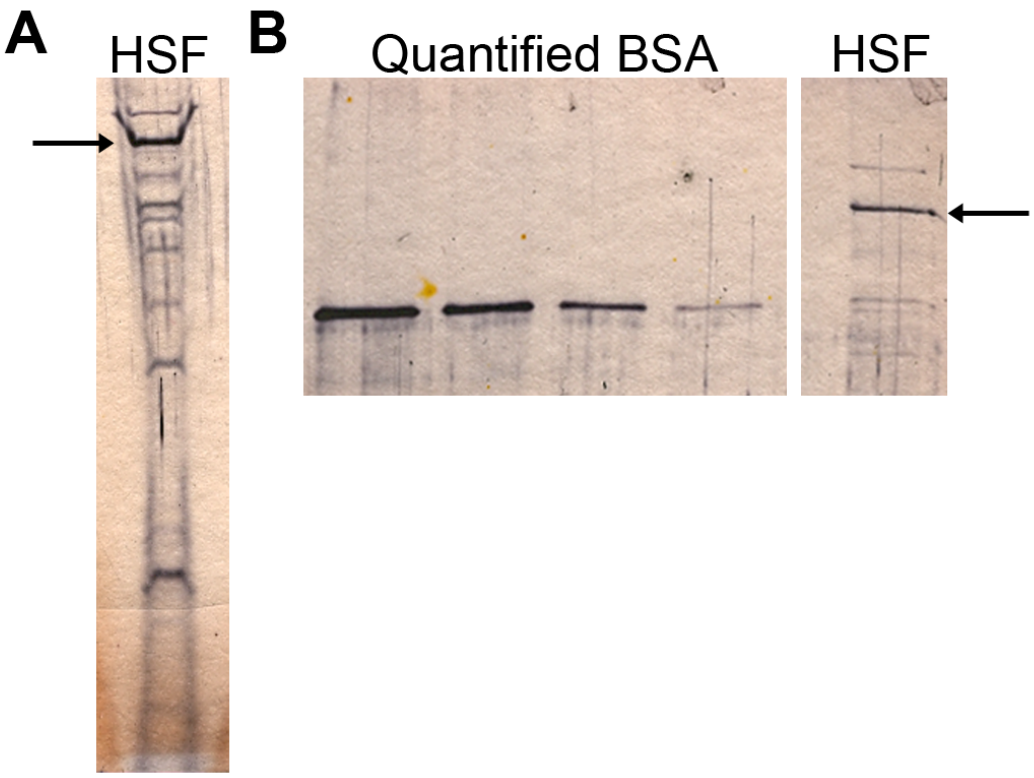

Figure A.1: HSF purification and quantification. A) Purified full-length HSF (arrow) was estimated to be 40% pure as quantified by a silver stained gel and densitometry. B) A silver stained gel using known concentrations of BSA (10 ng/l, 5 ng/µl, 2.5 ng/µl, 1.25 ng/µl) was used to quantify the stock concentration of purified full-length HSF (arrow) at 1.9 ng/µl. Note that one gel is shown, but intervening lanes were removed for simplicity.

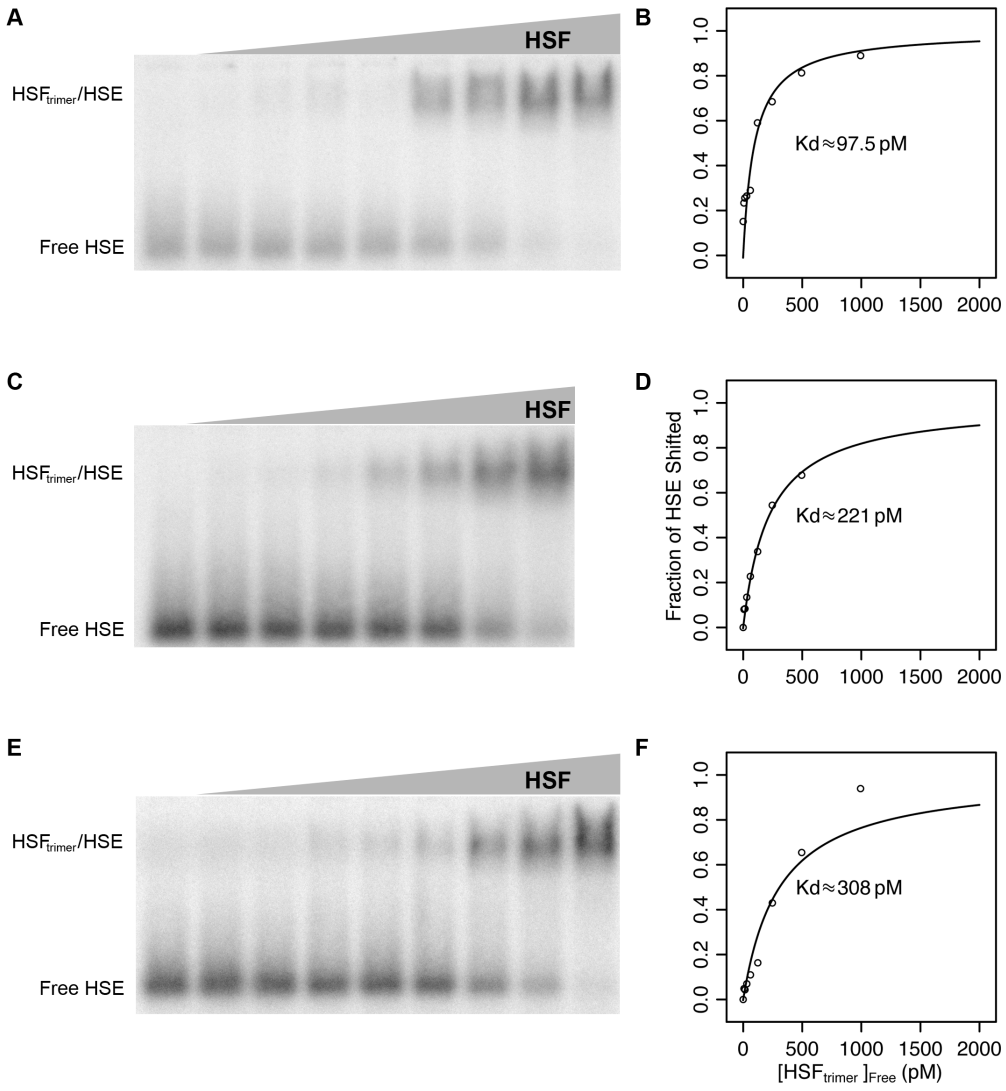

Figure A.2: A) The mobility of the constant 200 attomole HSE probe shifts into a trimeric-HSF:HSE complex as increasing HSF is added. There is no HSF in the left-most lane, the right-most lane contains 3 nM HSF (1 nM trimeric HSF), and the intervening lanes contain two-fold serial dilutions of HSF. B) A hyperbolic curve based on the Kd equation (see Methods) was modeled using the band shift data, indicating a Kd of 97.5 pM (95% confidence interval of 59.8-158 pM). C) The constant 200 attomole HSE probe shifts into a trimeric-HSF:HSE complex as increasing HSF is added. There is no HSF in the left-most lane, the right-most lane contains 1.5 nM HSF (500 pM trimeric HSF), and the intervening lanes contain two-fold serial dilutions of HSF. D) A hyperbolic curve based on the Kd equation (see Methods) was modeled using the band shift data, indicating a Kd of 221 pM (95% confidence interval of 197-250 pM). E) This panel has the same description as panel A. F) A hyperbolic curve based on the Kd equation (see Methods) was modeled using the band shift data, indicating a Kd of 308 pM (95% confidence interval of 214-448 pM).

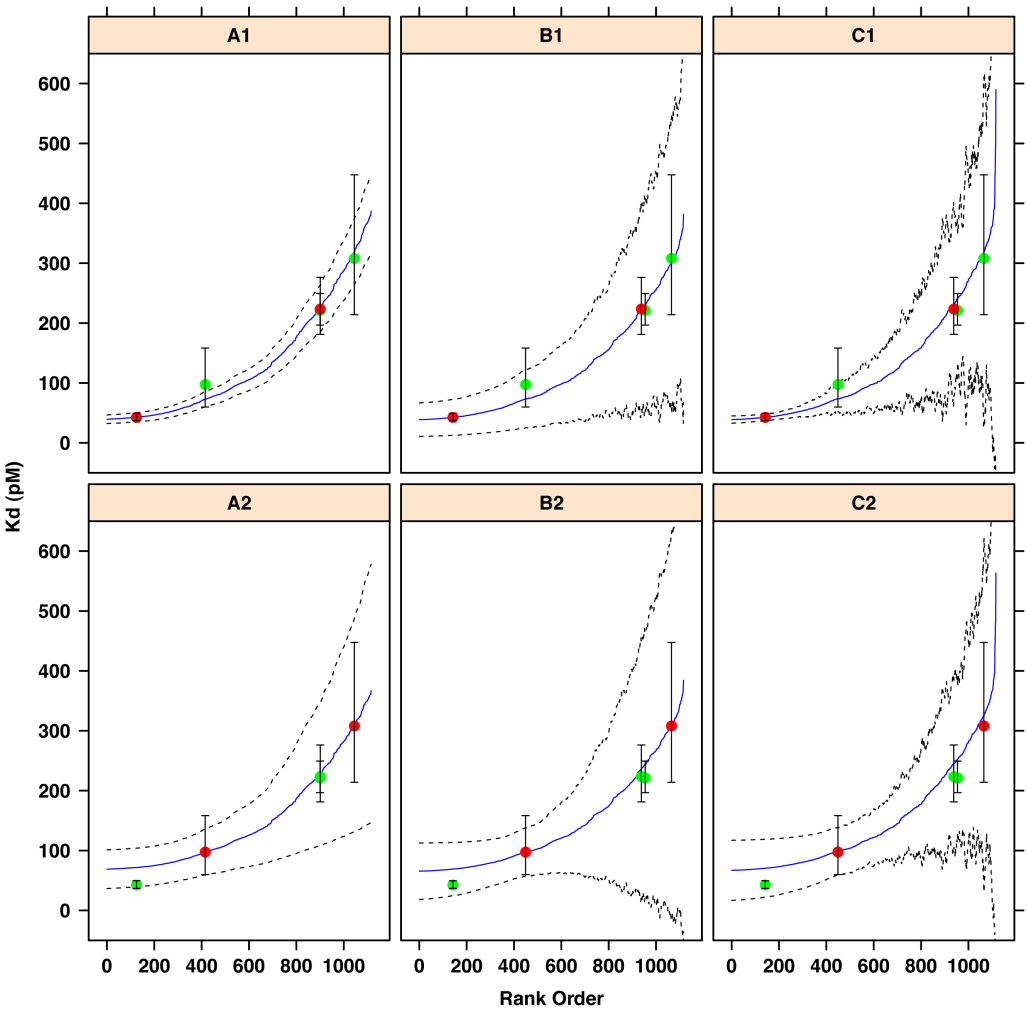

Figure A.3: Each panel shows smoothed 95% confidence intervals (CI) (dotted lines) for the estimated genomic Kd values (blue lines). The red and green points correspond to the Kd values determined by the EMSA assays. Error bars indicate 95% confidence intervals (CIs), as estimated in the non-linear regression (see Methods). Red points indicate those used as references to compute the genomic Kd values in each panel. The CIs shown in panels A1, B1, A2 and B2 were estimated by propagating various sources of uncertainty through our formula for estimating Kd values, using the first order Taylor expansion approximation. In panels A1 and A2, only the variance associated with the reference Kd points was considered, whereas in B1 and B2 the variance associated with the site intensity estimates was also used. At each binding site in the genome, the variance in intensity was estimated analytically from the two PB-seq replicates, after quantile normalization of the PB-seq replicate intensities to remove systematic biases. In panels C1 and C2, the CIs were computed by sampling the reference Kd values from normal distributions corresponding to their respective CIs and by selecting site intensities at random from one of the two PB-seq replicate values (again after quantile normalization). To account for the uncertainty associated with the choice of reference points, we show the CIs based on the two best EMSA points in the top panels and those based on the two worst EMSA points in the bottom panels.

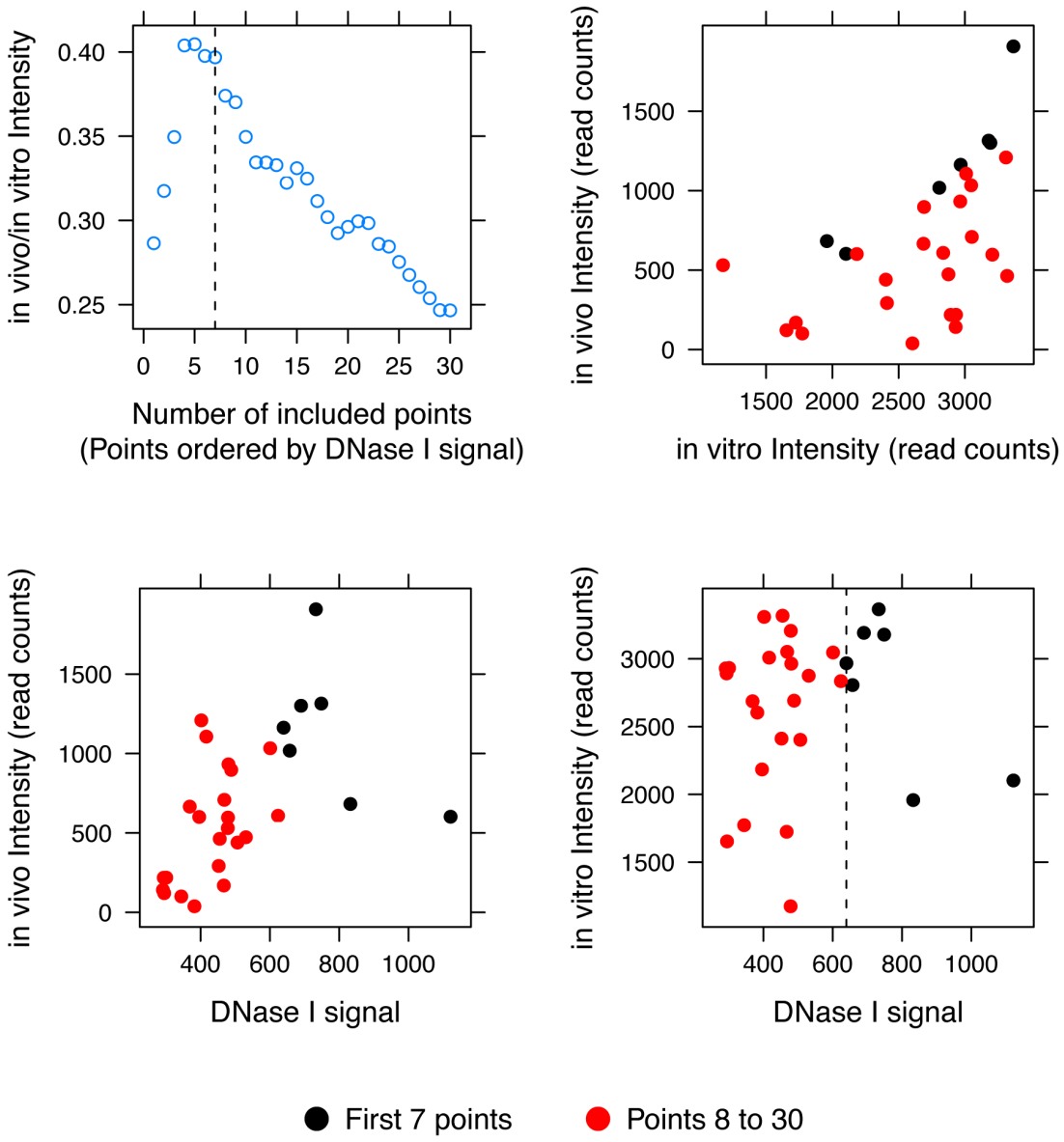

Figure A.4: These data points (HSE cluster sites) were used to determine the scaling factor between in vivo and in vitro binding intensities in Figure 2.1 and Figure 2.3. The top left plot shows how the in vivo to in vitro intensity ratio varies with the number of points included; dashed line signals the final choice of seven points. Scatter plots show the top 30 data points (HSE cluster sites) with the highest DNase I signal, against their in vivo and in vitro intensity values; black indicates the seven chosen points. The points with higher DNase I hypersensitivity offer the best choice for unbiased scaling.

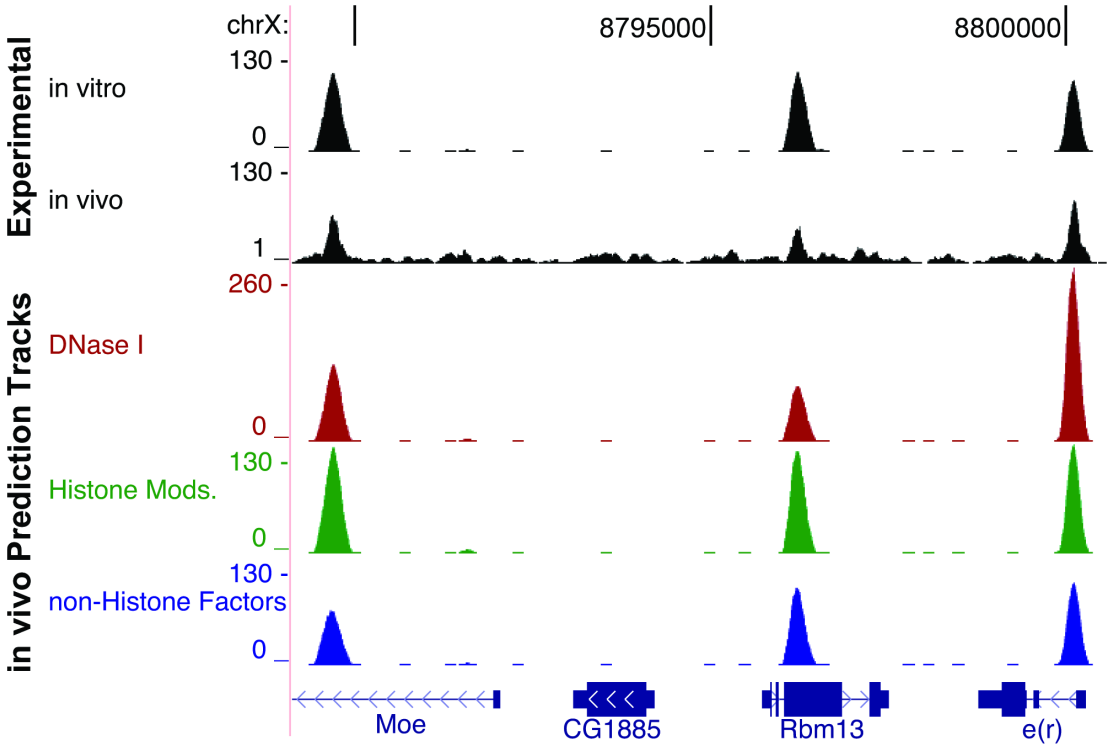

Figure A.5: This UCSC genome browser shot provides additional examples of in vivo prediction of HSF binding intensity using chromatin and PB-seq data.

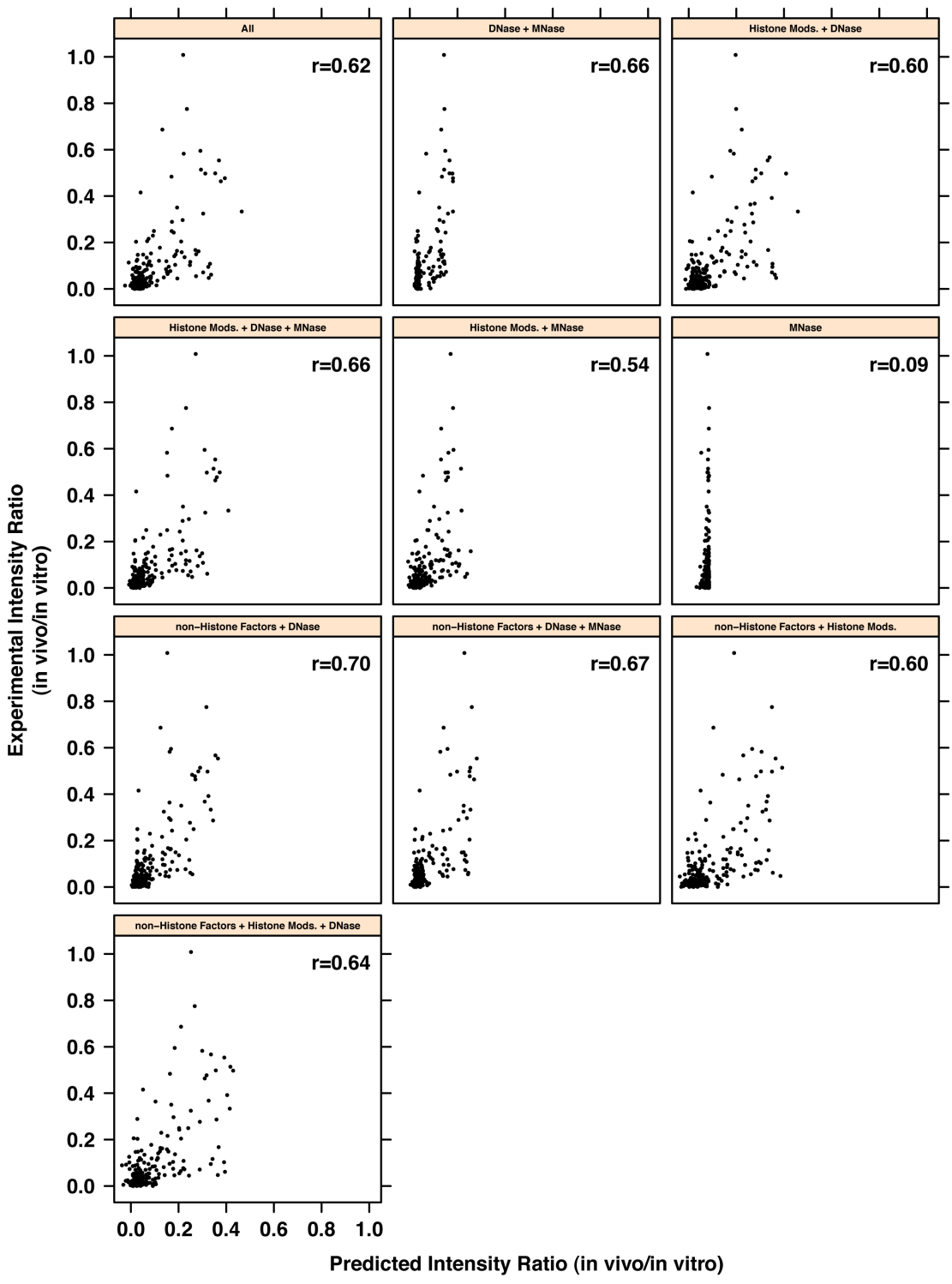

Figure A.6: The experimentally determined ratio between in vivo ChIP-seq HSF intensity and in vitro PB-seq intensity is plotted against the predicted in vivo/actual PB-seq ratio. The Pearson correlation for each model is shown.

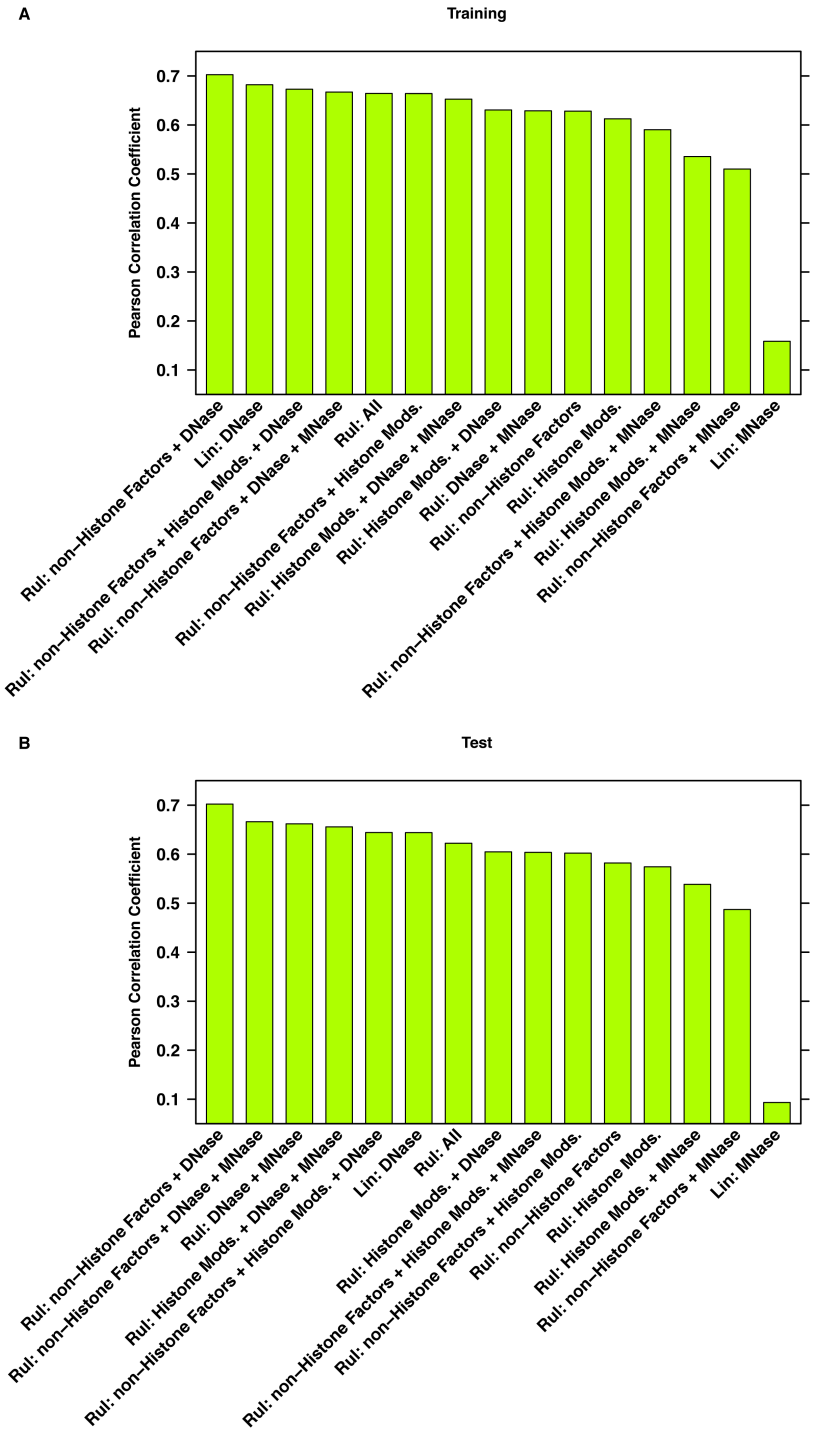

Figure A.7: The bar graphs indicate the Pearson correlation of predictions versus experimental measures for each model used to predict the in vivo/in vitro binding intensity ratio (Rul: Rules Ensemble model, Lin: linear regression model). The correlations for both the training data (panel A) and the test data (panel B) are indicated.

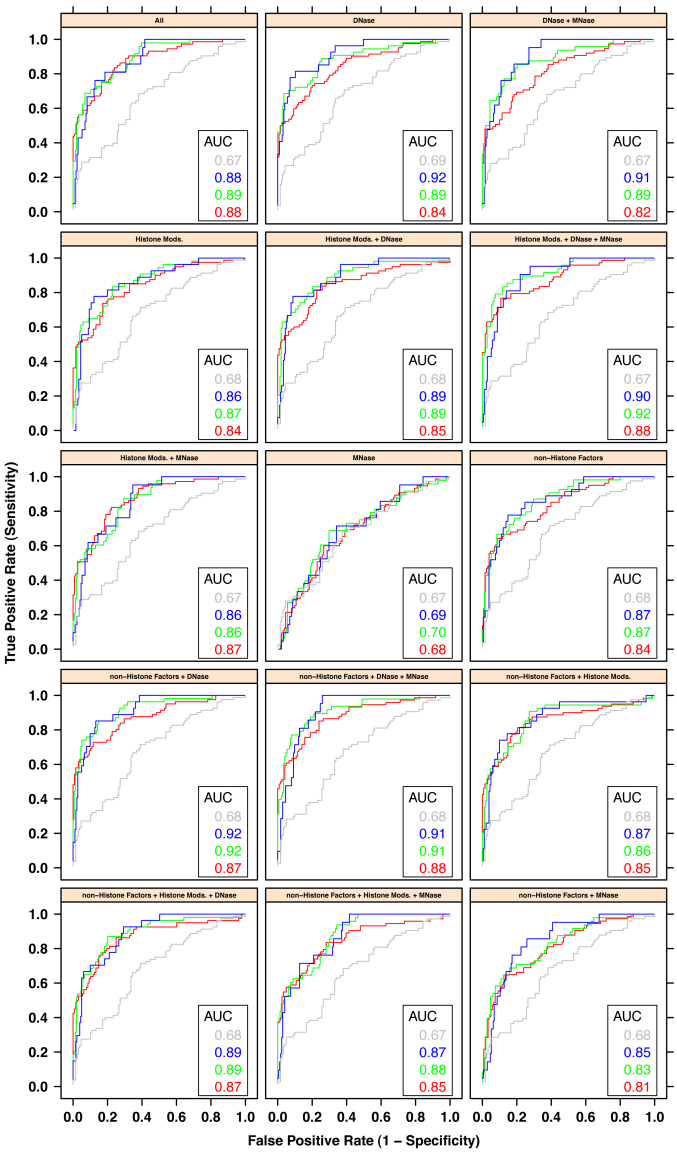

Figure A.8: ROC plots for in vivo HSF binding predictions. In vitro HSE sites were partitioned into bound and unbound cases by applying a threshold to the estimated in vivo intensity values. Three thresholds were considered: a permissive threshold (shown in red; 36% bound), a moderate threshold (green; 24% bound) and a strict threshold (blue; 12% bound). Each panel in the figure represents a distinct covariate set (see panel titles). For each covariate set, the corresponding rules ensemble model was applied to predict the in vivo intensity of the HSE sites. Each site was then classified as predicted to be bound or unbound by applying a threshold to these predicted intensities. These thresholds were varied to produce the Receiver Operating Characteristic (ROC) curves shown. As a baseline, we show predictions based on the scaled in vitro intensities in gray. For each ROC curve, we compute the Area Under the Curve (AUC) as a general measure of prediction performance (higher is better). Notice that the ROC curves are not highly sensitive to the threshold that is applied to the in vivo intensities, but in most cases the ensemble model produces a substantial improvement over the baseline prediction. At the same time, some covariates produce substantially better predictions than others.

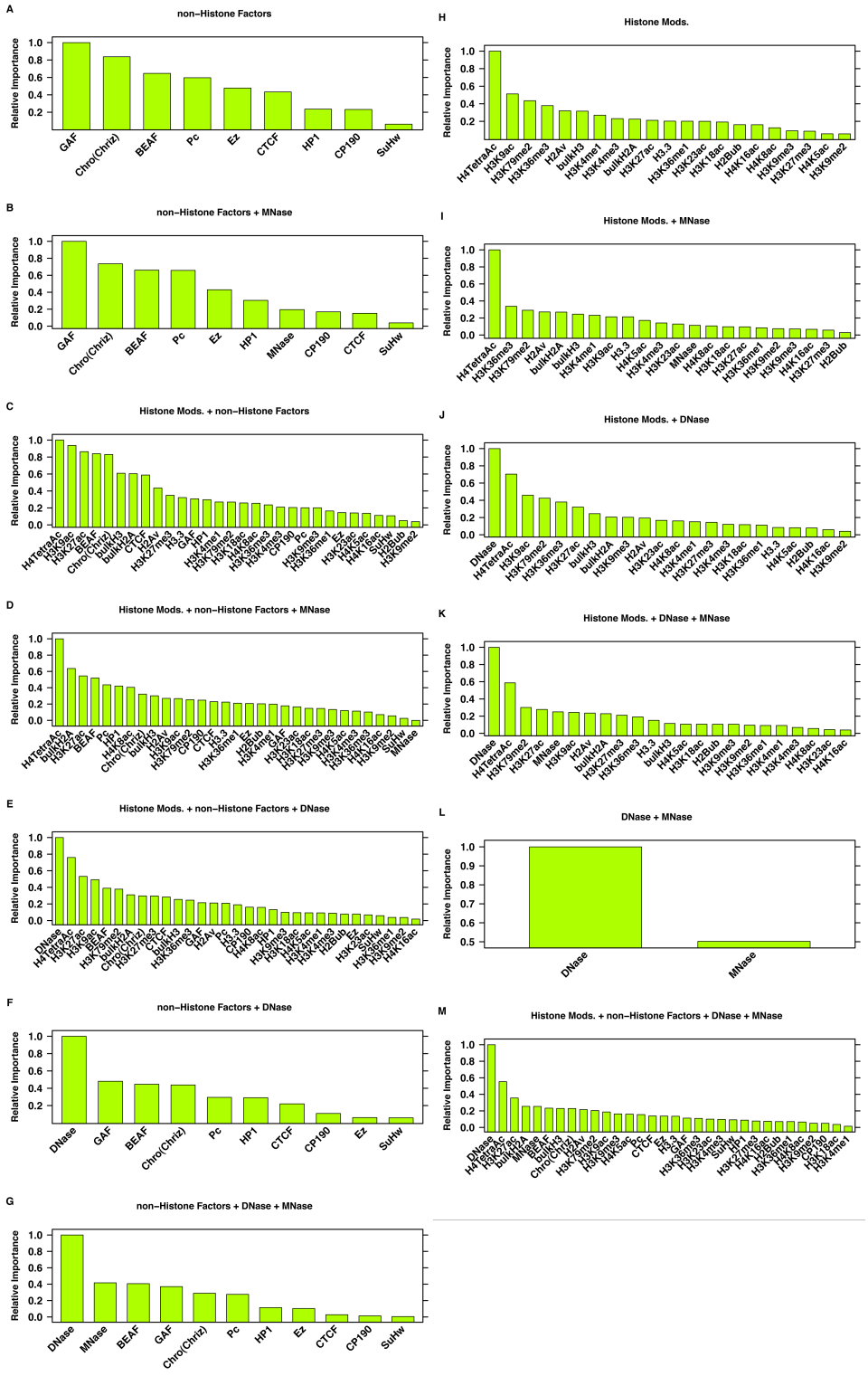

Figure A.9: For each of the models shown in Figure A.6 we show the relative importance [43] of each covariate in the Rule Ensemble model built with each indicated subset of covariates to predict the in vivo/in vitro binding intensity ratio.

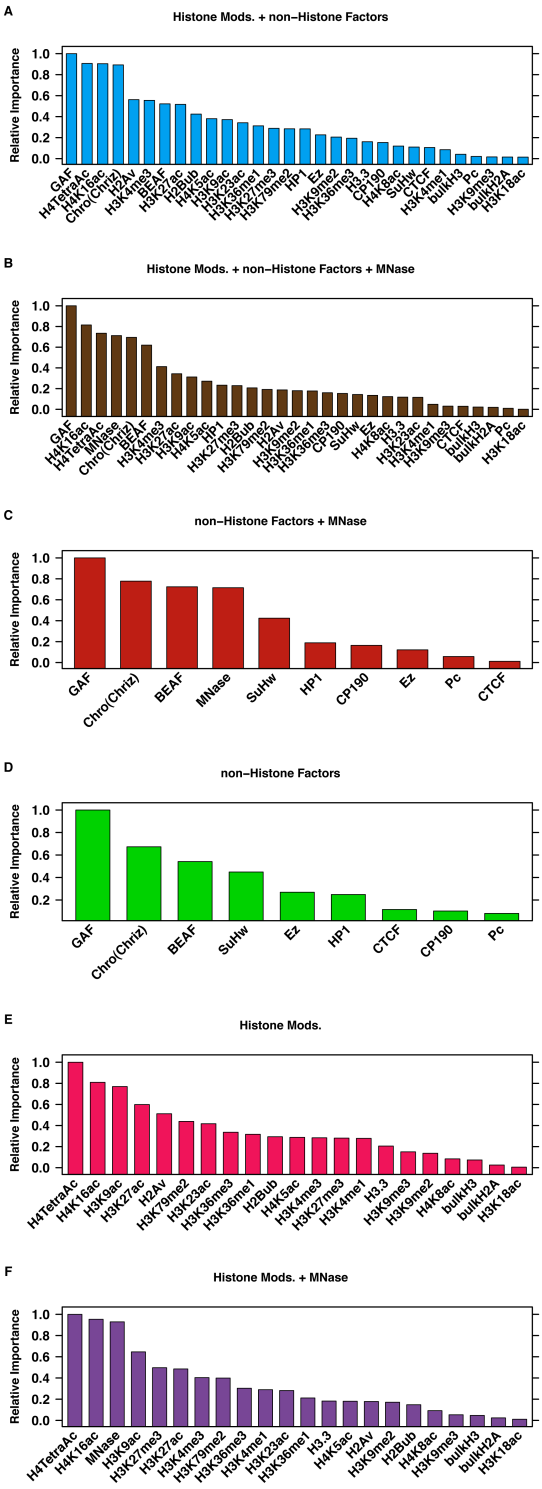

Figure A.10: The bar graphs illustrate the relative importance [43] of each covariate in the Rule Ensemble model built with each indicated subset of covariates to predict DNase I hypersensitivity.

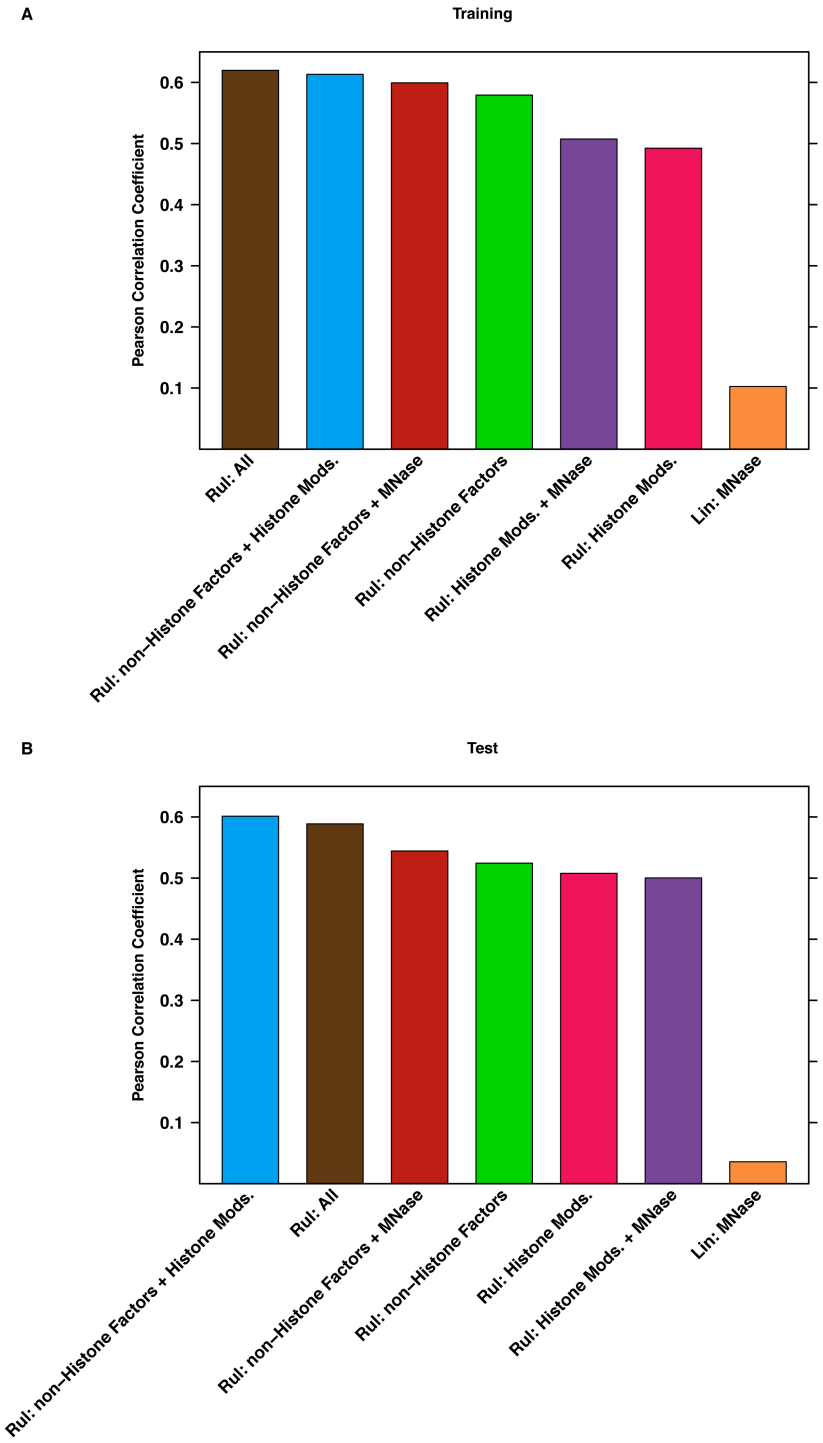

Figure A.11: These bar graphs indicate the Pearson correlation of predicted versus experimentally measured DNase I sensitivity for each DNase I prediction model (Rul: Rules Ensemble model, Lin: linear regression model). The correlations for the training data (panel A) and test data (panel B) are indicated.

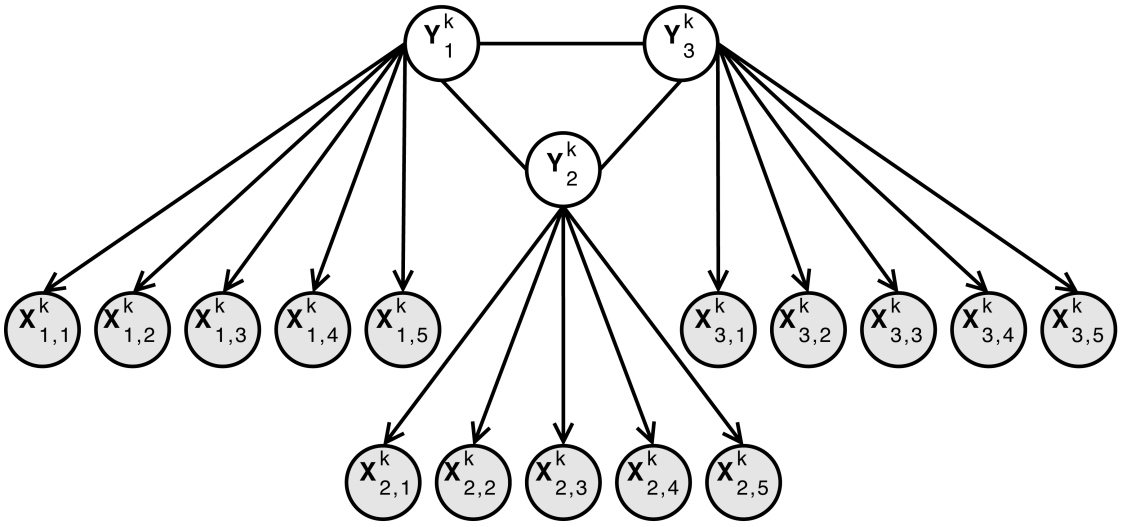

Figure A.12: The structure of the HSE probabilistic sequence model recapitulates the structure of the HSE. Each hidden variable Y_1,Y_2,Y_3, determines if the respective underlying pentamer bases are drawn from a strict base distribution or a relaxed version.

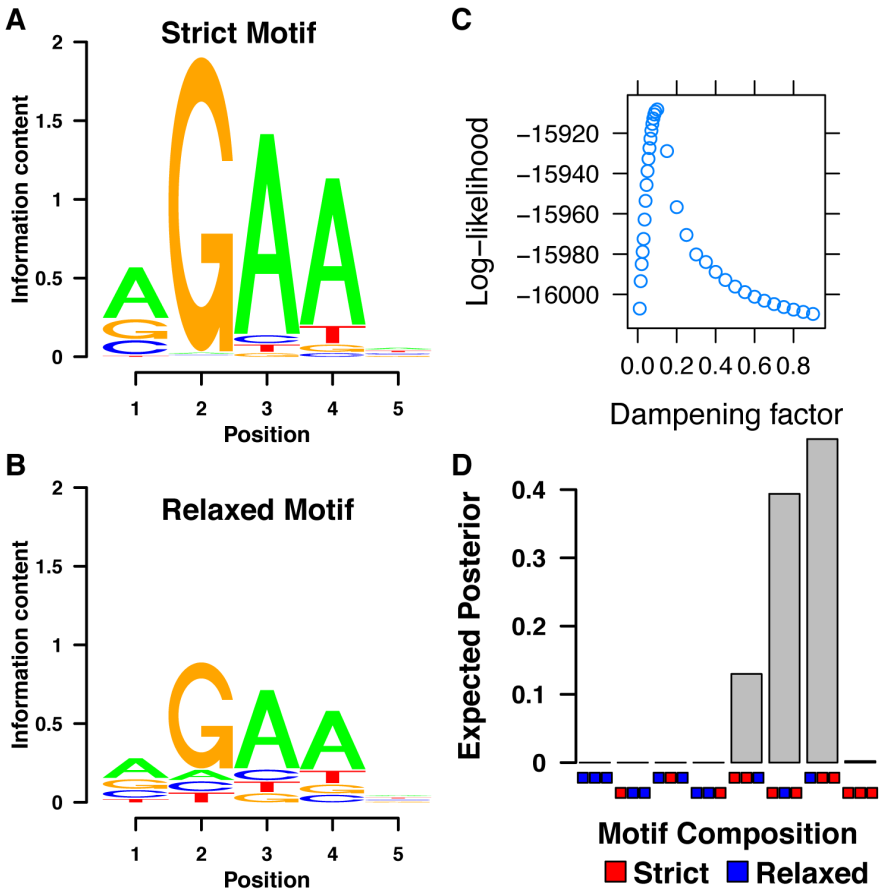

Figure A.13: Pentamers within the HSEs are dependent upon their stringency and position relative to the other pentamers. A) A composite pentamer matrix was derived from all pentamers found within PB-seq peaks. B) The strict motif from panel A and a dampening factor from panel C were used to generate a relaxed motif. C) The dampening factor was optimized to generate a relaxed motif that best explained the data. D) A probabilistic sequence model reveals that the presence of two strict and one relaxed pentamer provides the best explanation of the data.

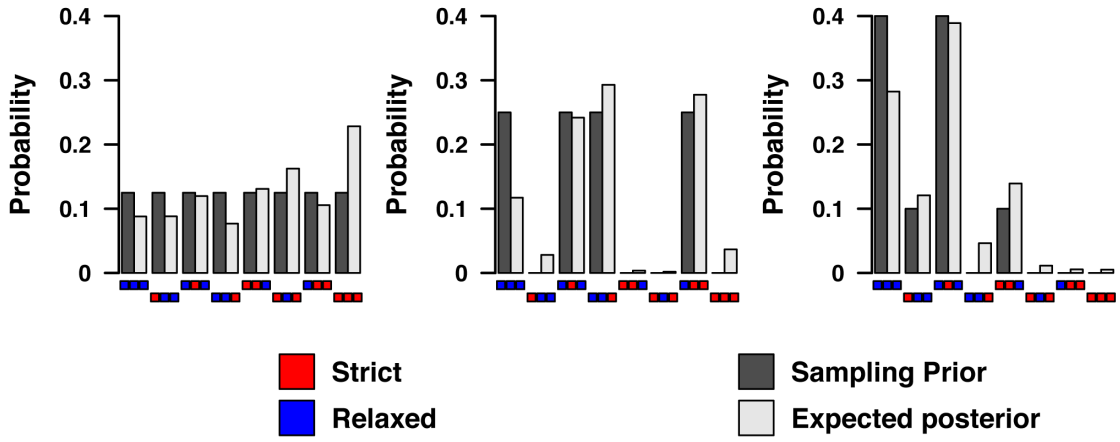

Figure A.14: The reduced HSE sequence model predictions are compared for patterns of strict/relaxed pentamer combinations. Three different simulated patterns are shown and are recapitulated by the model.

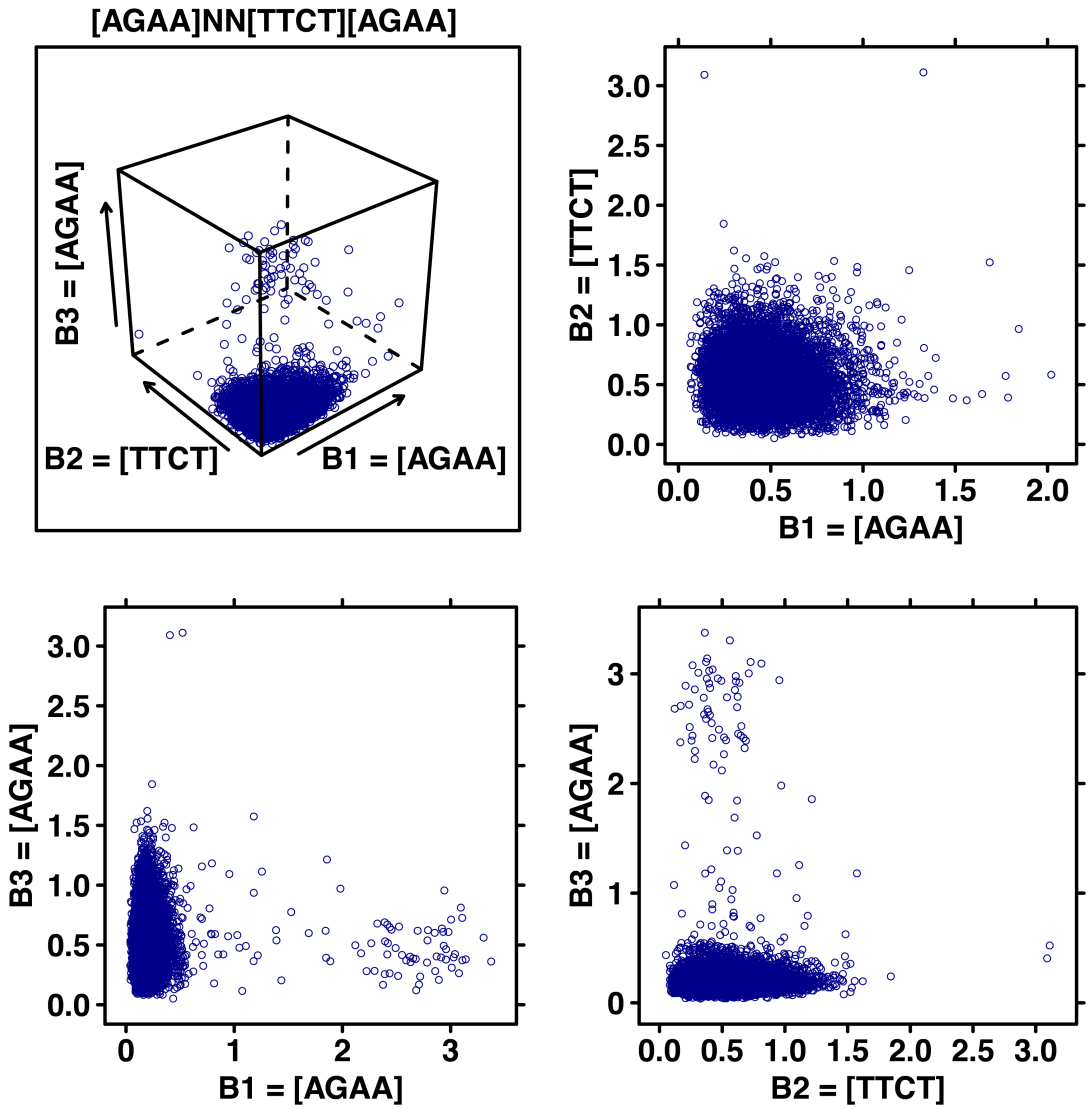

Figure A.15: Scatter plots show similarity of each HSE pentamer to the canonical monomer PSSM. Each point represents a PSSM estimated via MEME by sub-sampling the in vitro peaks identified by MACS. Pattern of the scatter plot shows evidence for pentamer divergence occurring on one pentamer at a time (points are spread following the axis, mainly corresponding relaxed versions of the first and second pentamers).

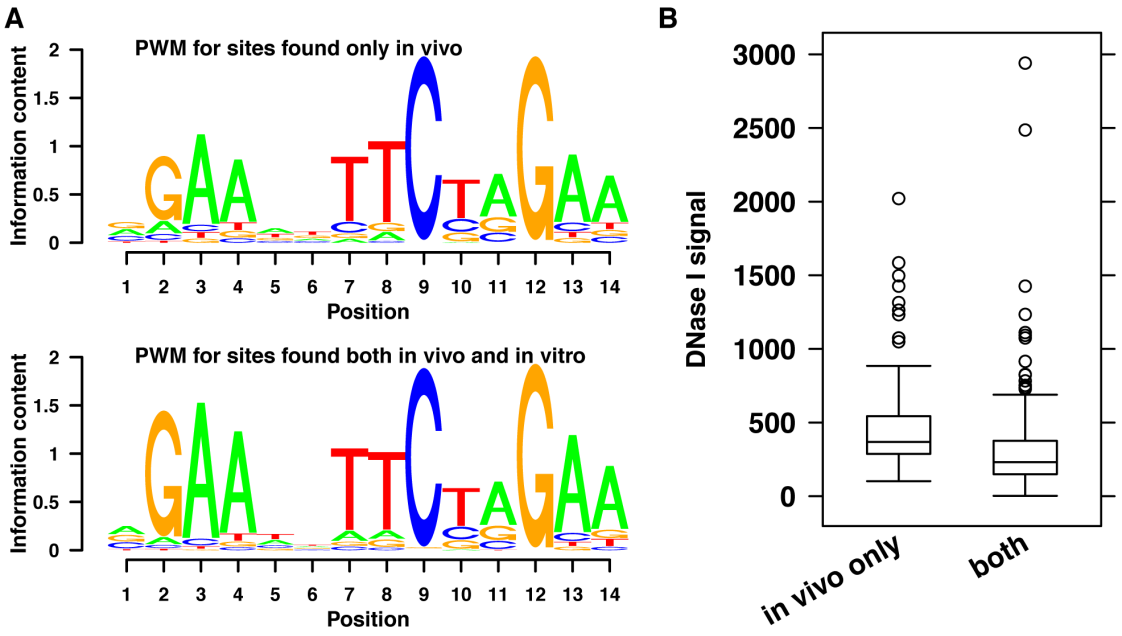

Figure A.16: In vivo HSF binding sites that were either detected or not detected in vitro have distinct properties. A) The composite PSSM for the 40% of HSF binding sites that are only found in vivo exhibits more degeneracy than the PSSM from the sites that are found both in vivo and in vitro. B) The binding sites exclusively found in vivo are generally more accessible, as measured by DNase I signal, than those sites found both in vivo and in vitro.

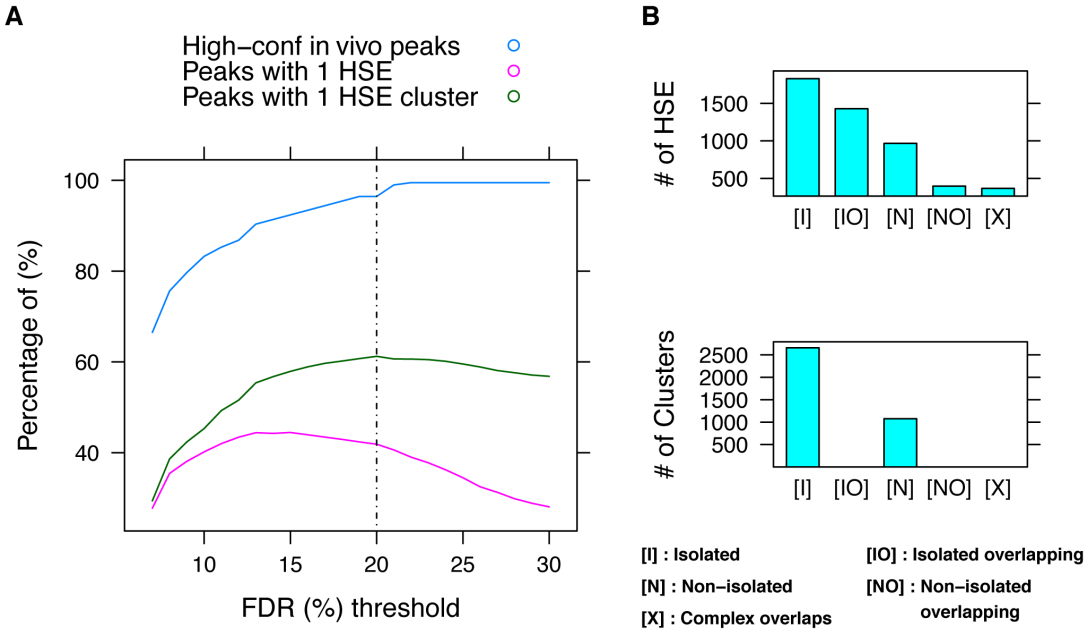

Figure A.17: A) Balance between in vivo recall and number of per peak in vitro HSE is reached at 20% estimated FDR, corresponding to the inflection point for the number of clusters, as well as near maximal recall of high-confidence in vivo sites. B) An HSE (or HSE cluster) is considered isolated if the nearest neighbor is more than 200 bp away. An HSE (or HSE cluster) is considered overlapping if it overlaps with a single other HSE (or HSE cluster); overlaps between more than two HSE (or HSE clusters) are denoted as complex overlaps.

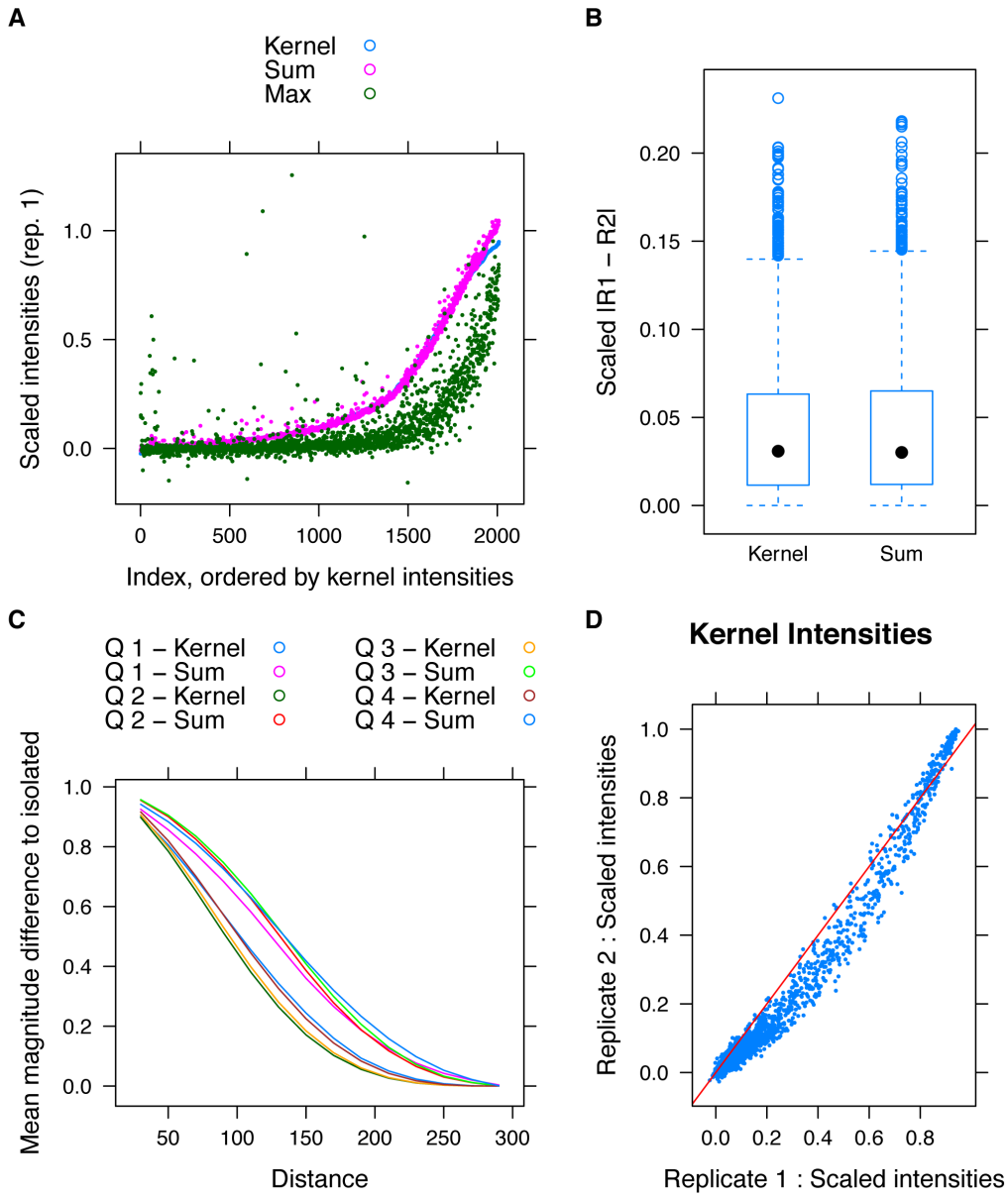

Figure A.18: Three different measures were compared for computing HSE cluster intensities: max, sum and bi-weight kernel. A) Each measure was incorporated into a scatter plot of scaled intensities. Max was rejected because it produced a more compressed range of intensity values. B) In comparing the difference in intensities across replicates, the bi-weight kernel approach fares slightly better than the sum. C) The difference in magnitudes given a cluster distance on isolated clusters was compared between measures. For each distance, the isolated clusters are made to overlap an identical copy of themselves and the magnitude difference is computed by comparing the value of the isolated cluster with the partially overlapping, using either the sum or kernel measures. Average values per intensity quartile show that bi-weight kernel measure introduces less error as a function of distance than the sum measure. D) Replicate intensities strongly correlate, as predicted, using the kernel measure.

Table A.1: ModENCODE identification number or GEO accession number for each data set used in the paper.

Data Sources	
Factor/Modification	**modENCODE ID/GEO #**
Tetra-Ac H4	modENCODE_201
BEAF	modENCODE_274
Chriz	modENCODE_278
CP190	modENCODE_280
CTCF	modENCODE_283
Ez	modENCODE_284
GAF	modENCODE_285
H2B Ubiq	modENCODE_290
H3K18ac	modENCODE_292
H3K23ac	modENCODE_294
H3K27ac	modENCODE_296
H3K27me3	modENCODE_298
H3K36me1	modENCODE_3170
H3K36me3	modENCODE_303
H3K4me1	modENCODE_304
H3K4me3	modENCODE_305
H3K79me2	modENCODE_307
H3K9ac	modENCODE_309
H3K9me2	modENCODE_311
H3K9me3	modENCODE_313
H4K16ac	modENCODE_319
H4K5ac	modENCODE_321
H4K8ac	modENCODE_322
HP1	modENCODE_323
Pc	modENCODE_326
Su(Hw)	modENCODE_330
H2A.v	GSM333840
H3.3	GSM333871
H3	GSM333834
H2A	GSM391380
Dnase I	SRP010823
Mnase	GSM550123

APPENDIX B

SUPPLEMENTAL MATERIAL FOR CHAPTER 3

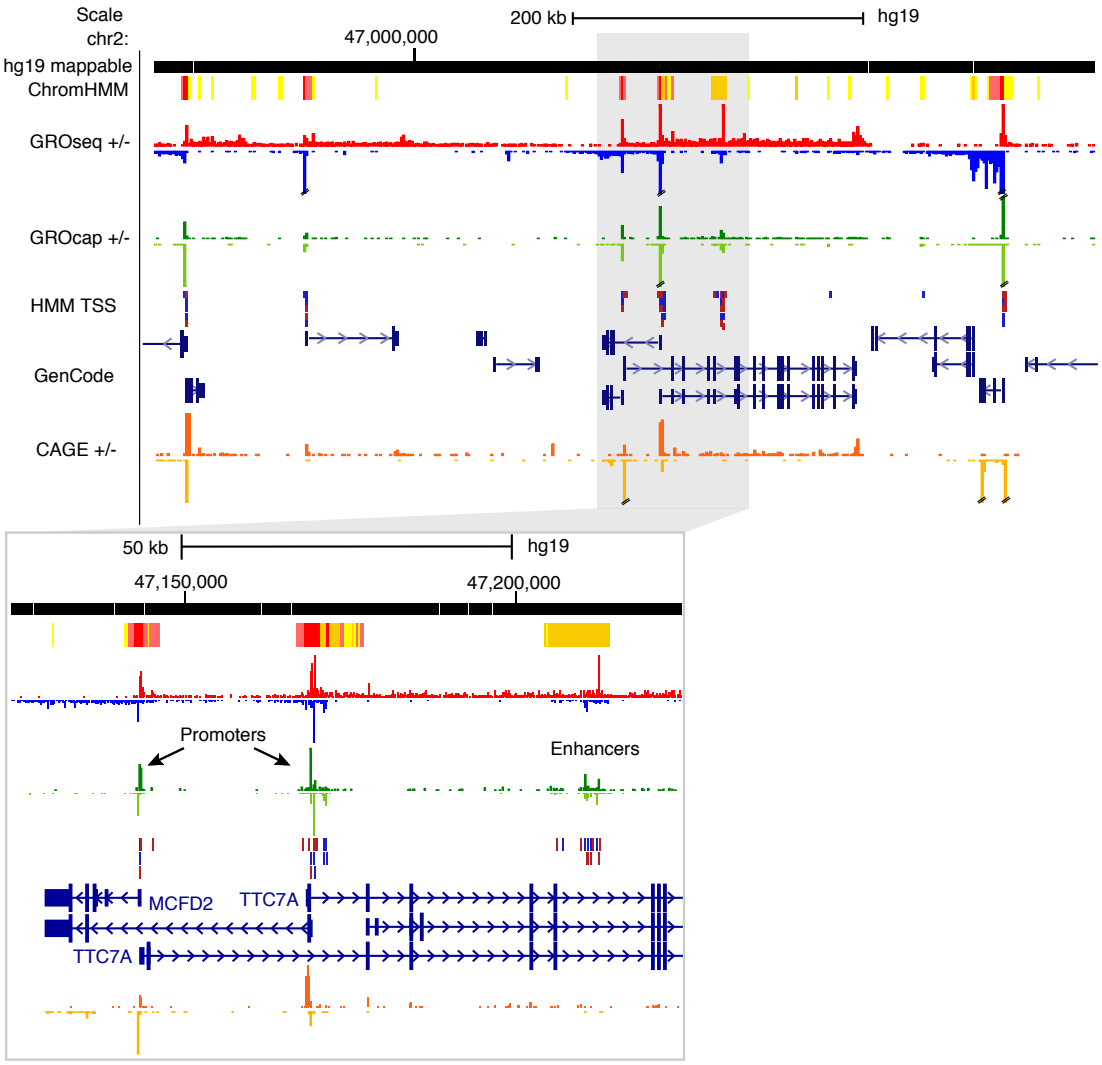

Figure B.1: Comparison of GRO-cap with CAGE. Shows a browser shot from the UCSC genome browser showing some of the data sets generated in this study (or previously published). The insert is a zoomed in view of the shaded region that shows the divergent GRO-cap (+ strand: dark green, - strand: light green) signal at a couple promoters (ChromHMM: red) and enhancers (ChromHMM: orange). Note that CAGE signal (+ strand: orange, - strand: light orange) is at background levels in the enhancer region. Data is from GM12878 cells.

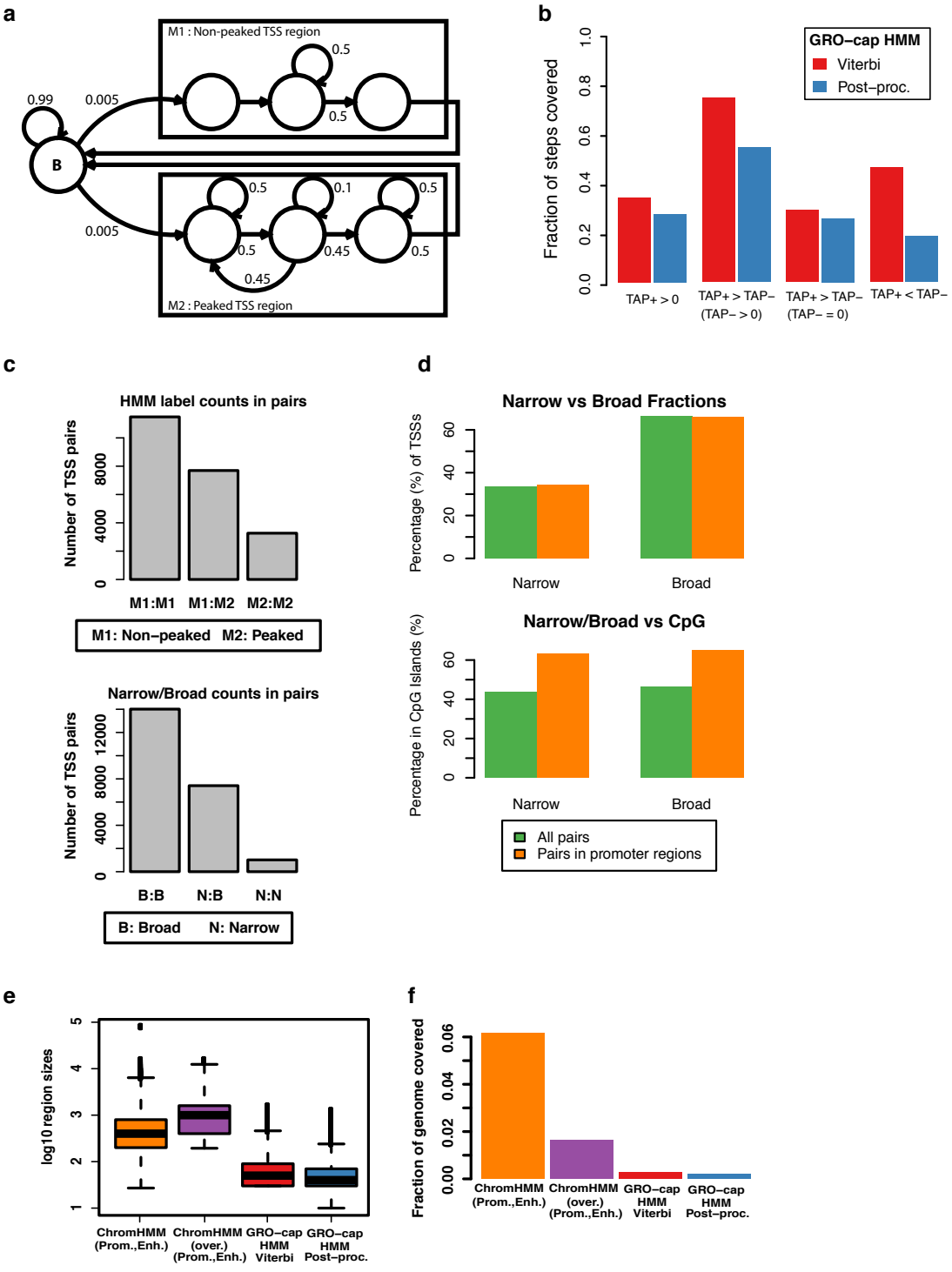

Figure B.2: (Caption next page.)

Figure B.2: (Previous page.) TSS Identification. (a) Hidden Markov model state diagram, with three state groups: B) background; M1) non peaked TSS region; M2) peaked TSS region (non-trivial transition probabilities indicated as labels to arrows). (b) Effect of TSS region prediction post-processing (see Methods) on coverage of GRO-cap data split by relationship between pre and post TAP signal. Overall, there is a small reduction in the fraction of the GRO-cap library that is covered (in number of 10 bp steps), with the largest reduction falling on depleted steps (TAP+ <TAP-). (c) Number of TSS pairs that correspond to each combination of peaked (M2) and non-peaked (M1) subsets (top) and each combination of broad (B) and narrow (N). (d) Narrow/broad distinction based on whether over more/less than 50% of GROcap reads are within +/- 2bp of the mode (best site). There is 45% agreement between the two ways to label pairs (assuming M1 = B and M2 = N). Comparison of narrow and broad TSS regions (from paired subset) with promoter annotations (ChromHMM, top panel) and CpG Island overlap (bottom panel; CpG Island track from UCSC Genome Browser). Narrow/broad distinction based on whether over more/less than 50% of GROcap reads are within +/- 2bp of the mode (best site). No significant difference is observed in either case. (e) Distribution of ChromHMM Promoter and Enhancer regions and GRO-cap TSS prediction lengths. For ChromHMM, we show both the full set (orange) and the subset that has an overlapping GRO-cap TSS prediction (purple). GRO-cap TSS prediction lengths are shown for both before (Viterbi, red) and after post-processing (blue) (see Methods). (f) Fraction of the genome covered by predicted TSS regions compared with ChromHMM Promoter and Enhancer regions. For ChromHMM, we show both the full set (orange) and the subset that has an overlapping GRO-cap TSS prediction (purple). GRO-cap TSS prediction lengths are shown for both before (Viterbi, red) and after post-processing (blue). Data are from GM12878 cells.

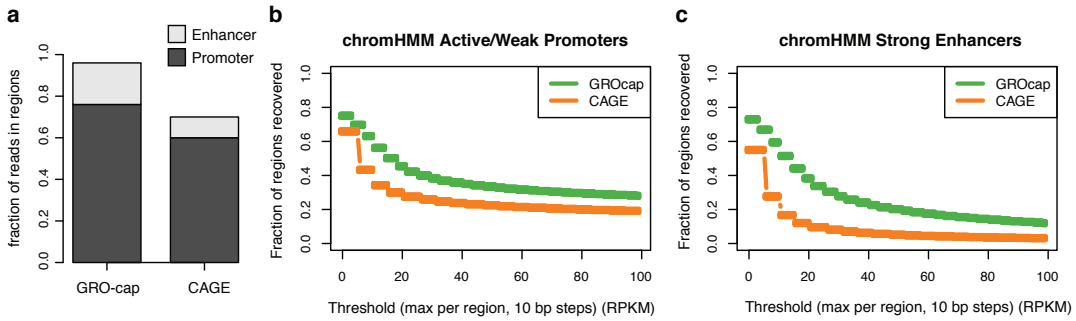

Figure B.3: Comparison of GRO-cap and CAGE. (a) Fraction of reads in promoters (light grey) and enhancers (dark grey) for GRO-cap (76% promoter, 20% enhancer) and CAGE (60% promoter, 10% enhancer). (b,c) Recovery threshold plots showing the fraction of total promoters (b), and enhancers (c) that are recovered at varying thresholds of GRO-cap (green) and CAGE (orange). Data are from GM12878 cells.

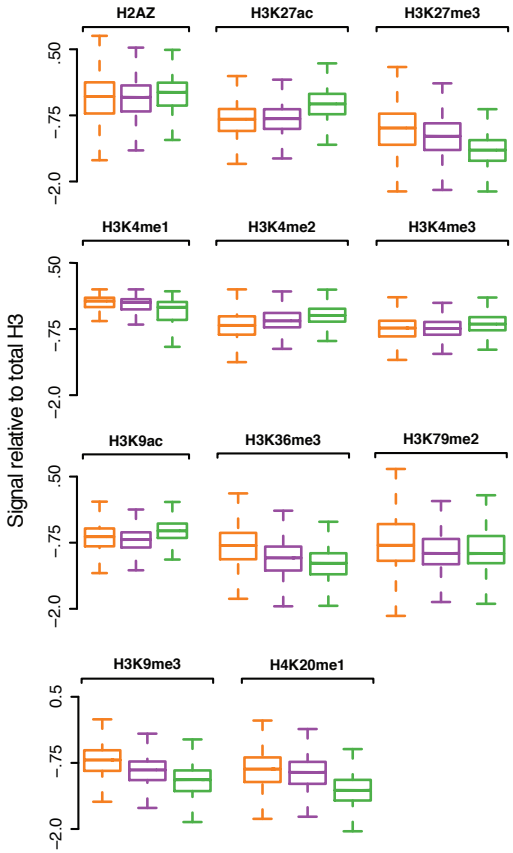

Figure B.4: Histone modifications in enhancer classes. Distribution of ChIP-seq histone modification signals in each enhancer class (poised: orange, open: purple, transcribed: green), scaled by total H3K4 methylation signal.

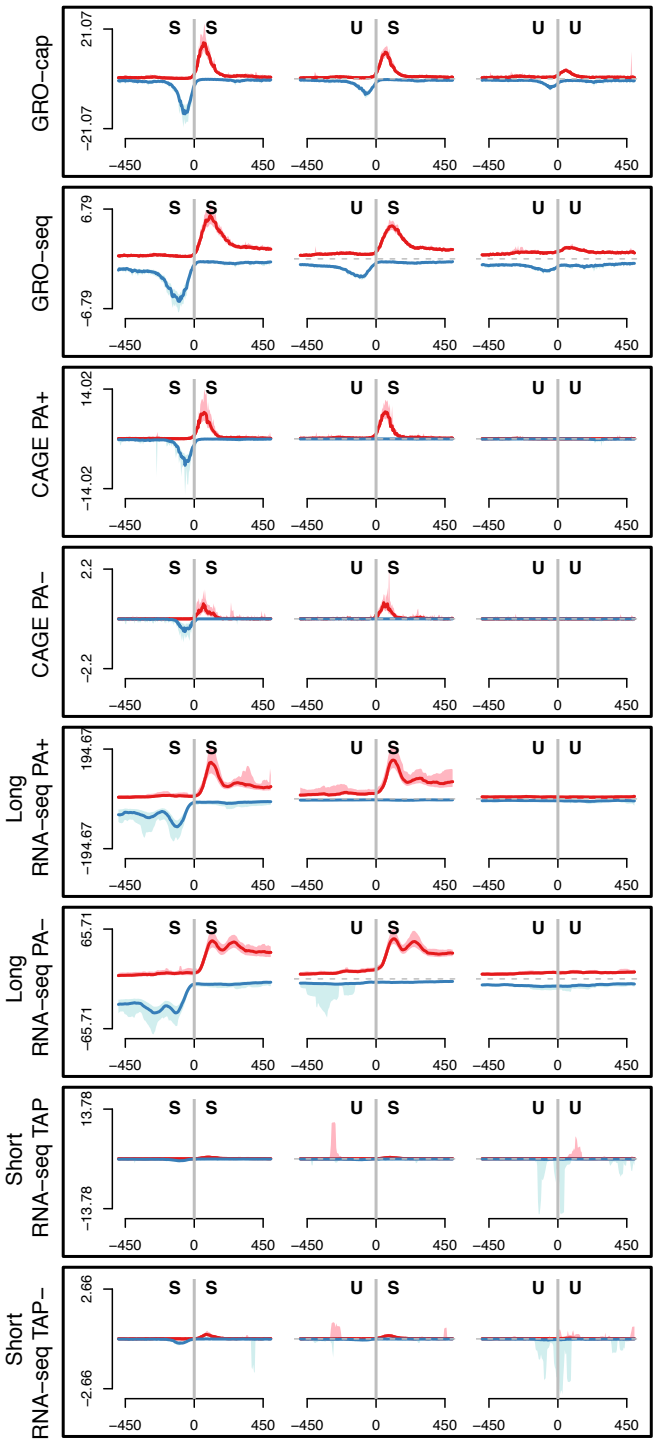

Figure B.5: Profiles of various RNA sequencing data at TSS pairs after stability classification. Metaplot profiles of various types of RNA sequencing data aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable (left), unstable::stable (center), unstable::unstabled (right). GRO-seq and GRO-cap data were produced for this study. All other data were produced by the ENCODE consortium [25]. Data are from GM12878 cells.

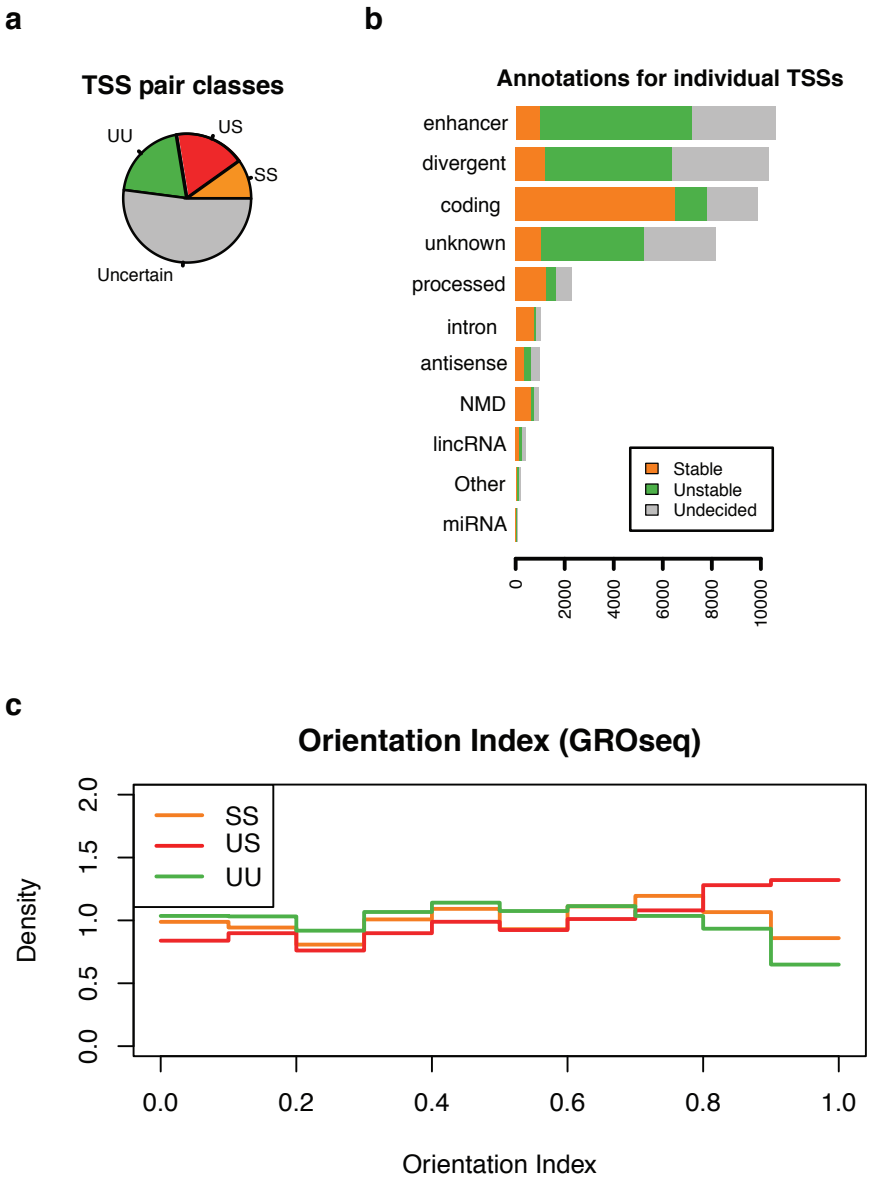

Figure B.6: TSS pair classes. (a) Pie chart shows relative proportion of TSS pair stability classes, including "Uncertain" for those in between the two thresholds. (b) Individual TSSs within pairs were matched to various annotations based on GENCODE annotations or ChromHMM regions (for enhancers). TSSs for each annotation were then split on stability classifications: stable (orange), unstable (green), undecided (gray). (c) Orientation indexes (OI) are presented for pairs classified as stable::stable (red), unstable::stable (orange), unstable::unstable (green). OI scales between zero (bi-directional) and one (uni-directional) and is defined as $2 \times (\max(Rp, Rm)/(Rp + Rm) - 1$. Rp and Rm are the plus and minus strand, respectively, GRO-seq reads that fall in the 250 bp downstream of the strongest (highest read count) GRO-cap position in the TSS region in each pair. OI was calculated from GRO-seq data in GM12878 cells.

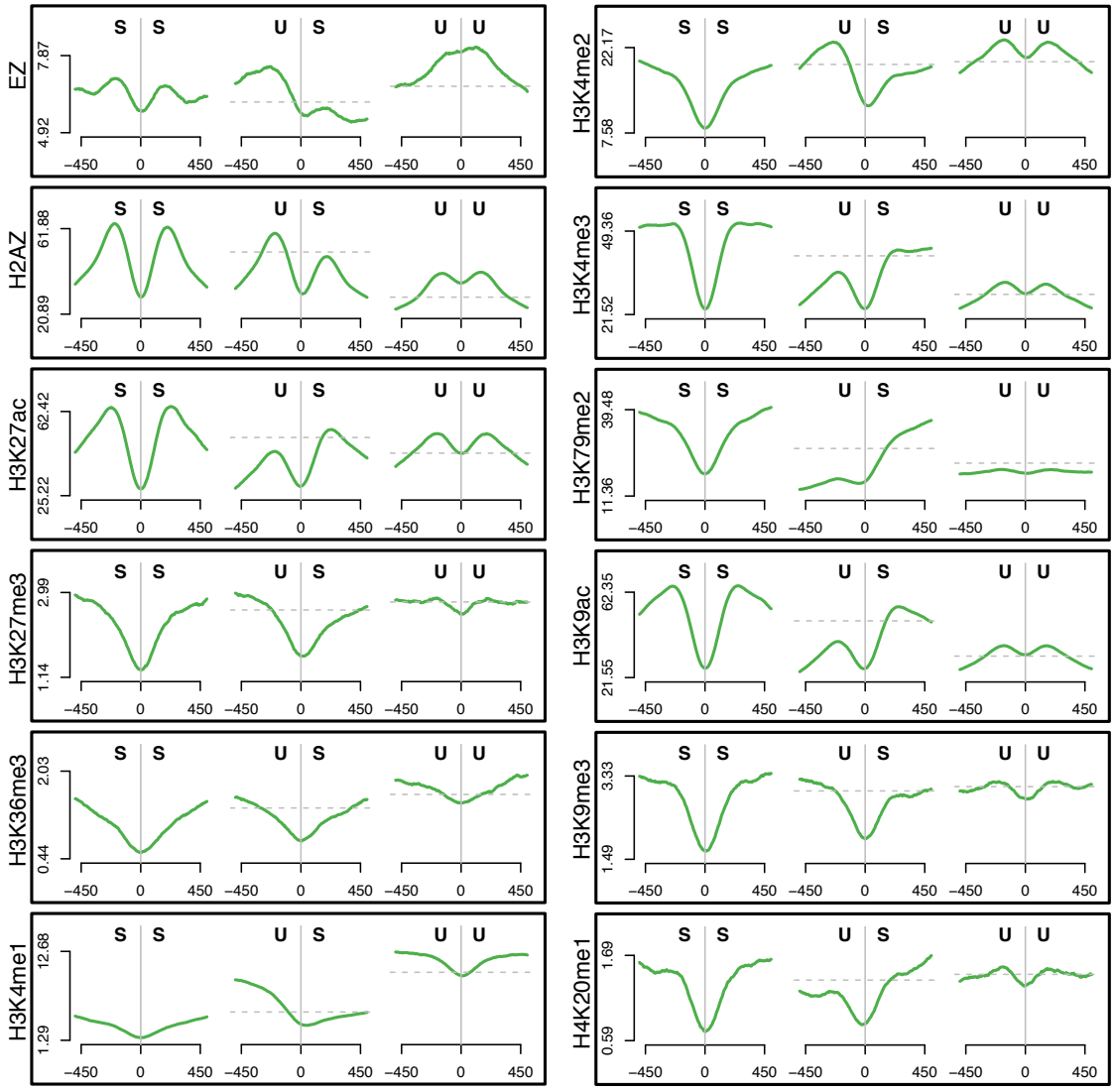

Figure B.7: Profiles of various histone marks or chromatin binders at TSS pairs after stability classification. Composite profiles of ChIP-seq data for various histone modifications, variants, or chromatin binding proteins aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable (left), unstable::stable (center), unstable::unstable (right). All ChIP-seq data were produced by the ENCODE consortium [25]. GRO-cap data were from GM12878 cells.

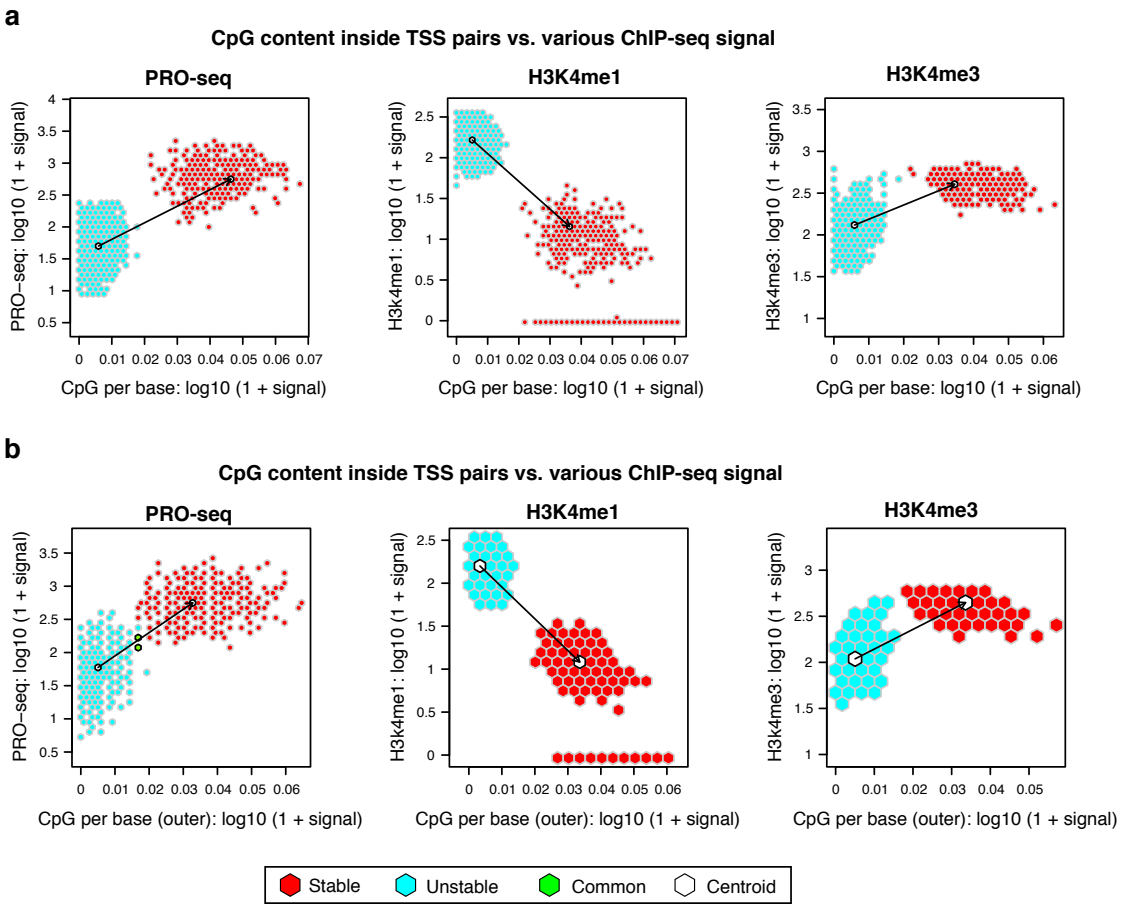

Figure B.8: CpG content vs. transcription and histone modifications at divergent TSSs. (a) CpG content inside TSS pairs versus PRO-seq signal (left), H3K4me1 (center) and H3K4me3 (right). Signal is further split between unstable (light blue) and stable (red) TSSs. Centroid for each subset in white. (b) CpG content outside TSS pairs versus PRO-seq signal (left), H3K4me1 (center) and H3K4me3 (right). Signal is further split between unstable (light blue) and stable (red) TSSs. Centroid for each subset is in white. PRO-seq data are from K562 cells.

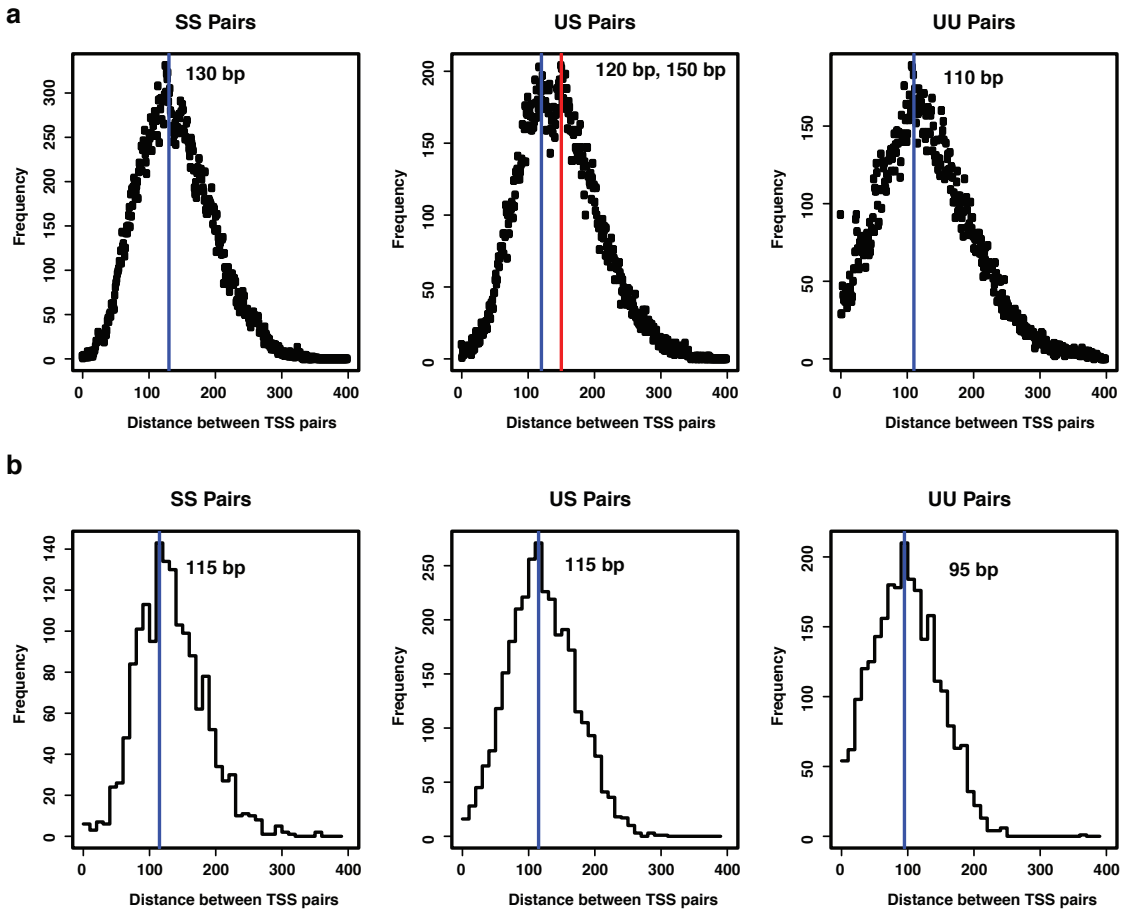

Figure B.9: TSS pair distances at TSS with different stability classifications. (a) Divergent TSS distance distribution obtained by computing, over TSS pairs, the distances between each combination of plus strand and minus strand positions within +/- 150bp of the divergent TSS center and where the GRO-cap signal is significantly above the control signal. Separate estimates obtained for each stability class (SS, US, UU). Estimates vary between 110 and 150 bp. (b) Divergent TSS distance distribution obtained by computing, over all TSS pairs, the distance between the centers of mass of the TSS regions in each strand. Estimates between 95 and 115 bp.

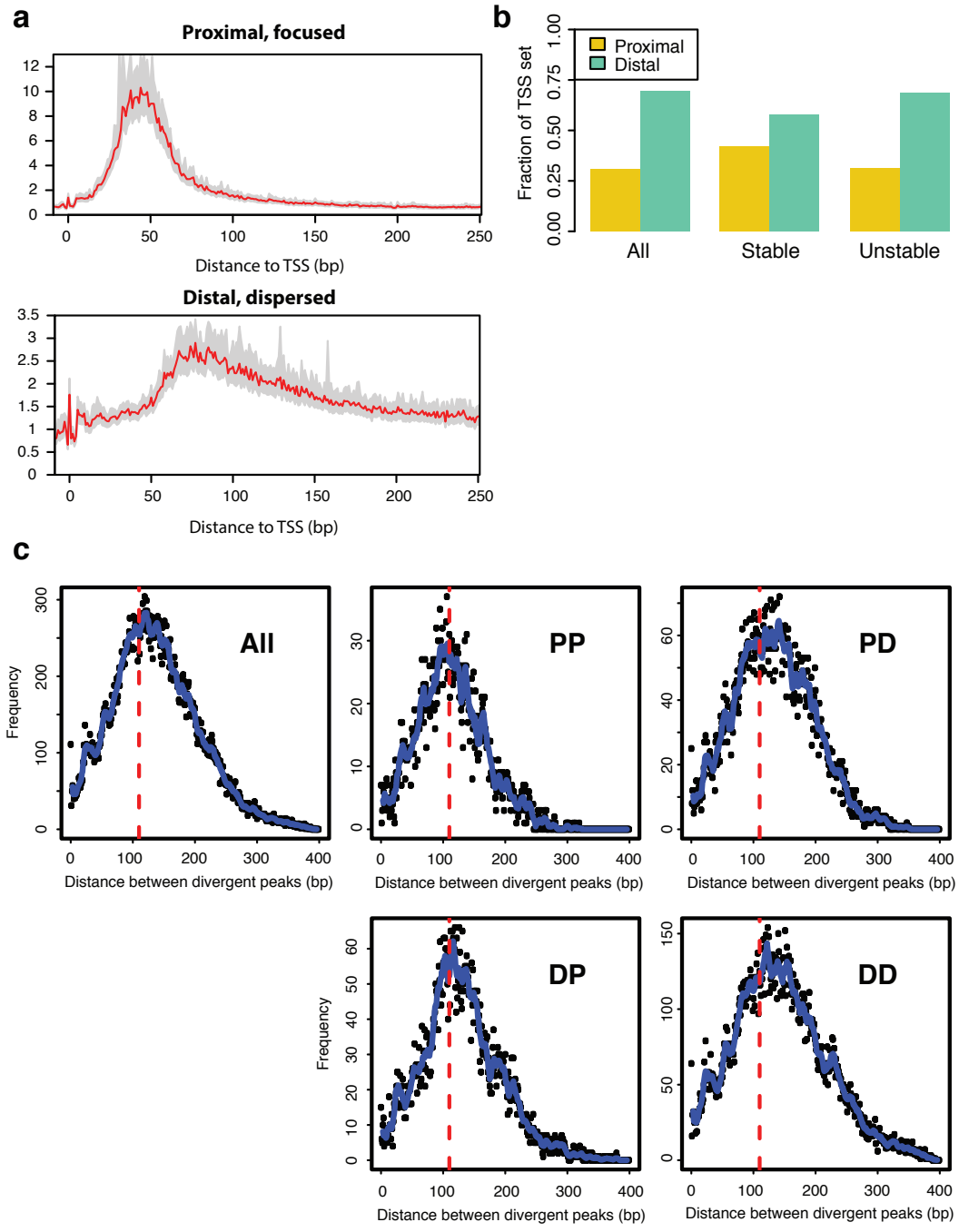

Figure B.10: Promoter-proximal pause versus TSS distance in pairs. (a) Two modes of promoter-proximal pausing are detectable (via k-means clustering): proximal-focused and distal-dispersed; this is consistent with previous results in *Drosophila* [81]. (b) Comparison of promoter-proximal pause modes with TSS stability classes shows an enrichment of distal-dispersed pause mode in unstable versus stable and an overall preference for distal-dispersed pausing across all TSSs. (c) TSS distances between divergent TSSs (in pairs) segregated by pause mode labeling on each side (P for proximal-focused, D for distal-dispersed). There is no apparent effect of pausing mode on distance.

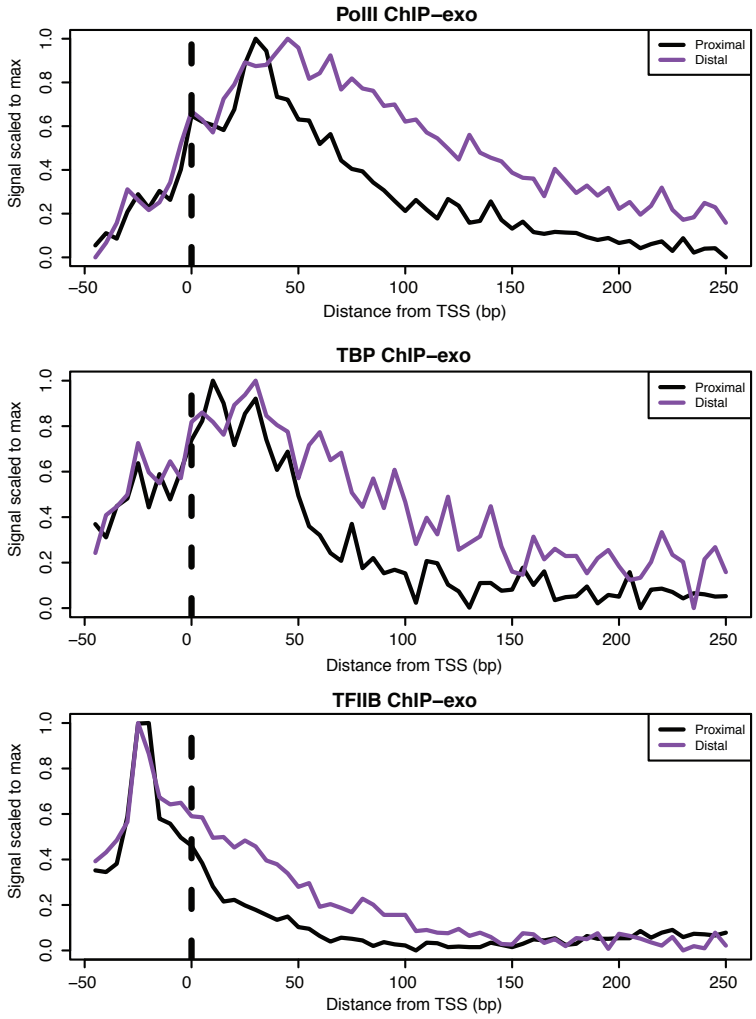

Figure B.11: Promoter-proximal pause versus core promoter factors. ChIP-exo data [138] composite plots of Pol II (top), TBP (middle) and TFIIB (bottom) aligned to GRO-cap TSS at both proximal-focused and distal-dispersed pause mode subsets. Note that ChIP-exo does not necessarily represent the position of each factor as they can cross-link to the DNA through other factors. Data are from K562 cells.

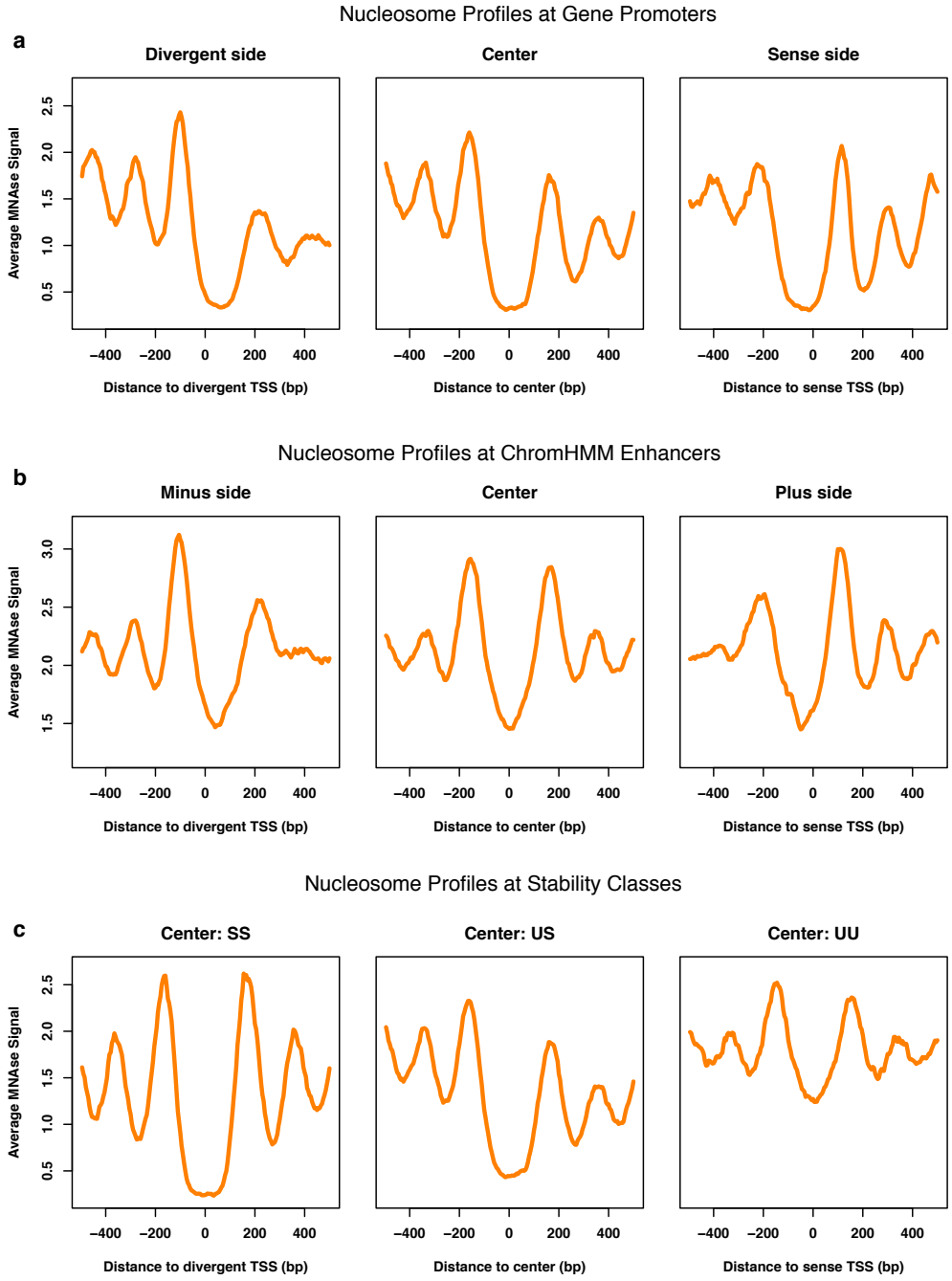

Figure B.12: Nucleosome profiles at TSS pairs. Nucleosome profiles at tSS pairs that map to (a) promoters and (b) enhancers. MNase-seq data is aligned to upstream TSS (left, divergent), the center of the pairs (center) or the downstream TSS within the pairs (right, sense). (c) Shows the nucleosome profiles aligned to the center of pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable (left), unstable::stable (center), unstable::unstable (right). MNase-seq data [25] and GRO-cap data from GM12878 cells.

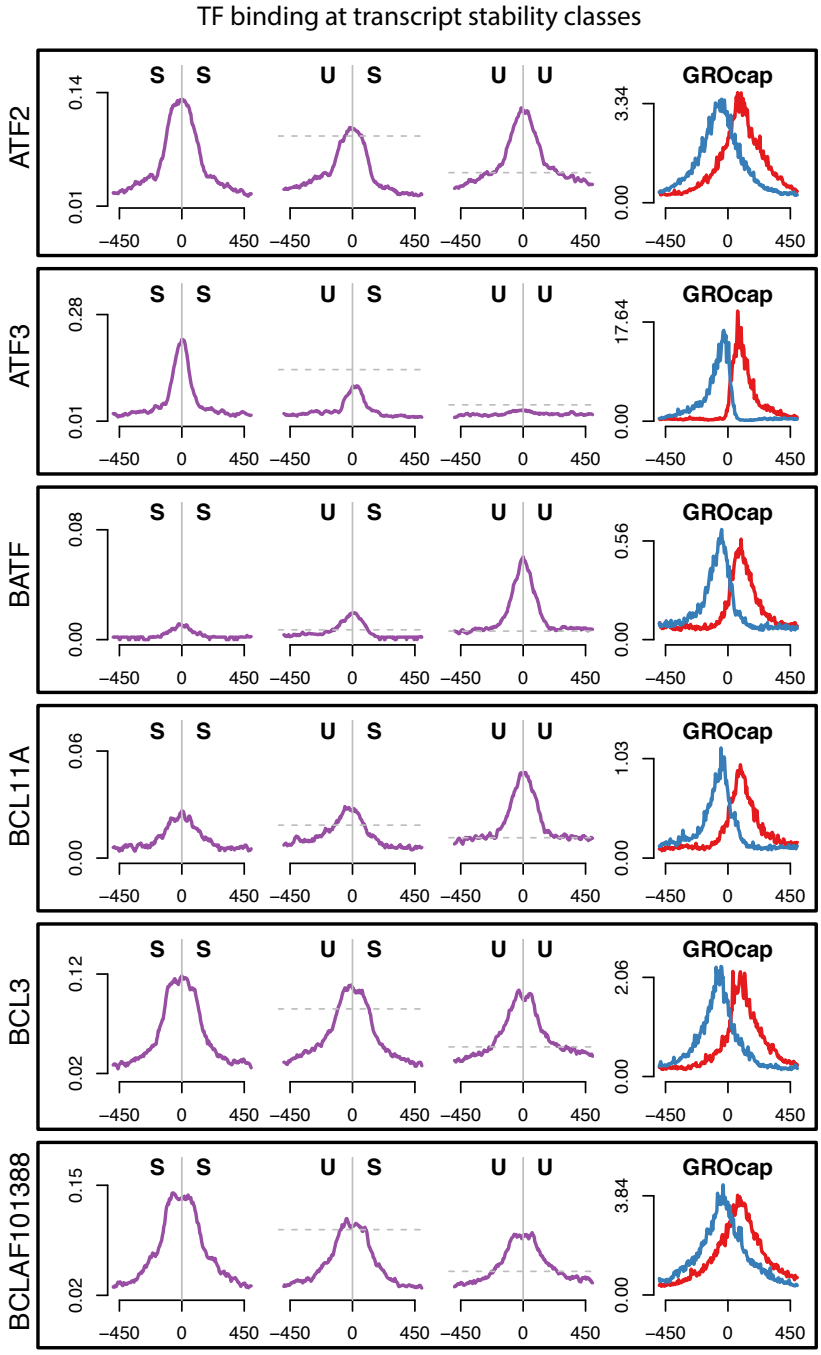

Figure B.13: Profiles of transcription factors at TSS pairs after stability classification (1/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

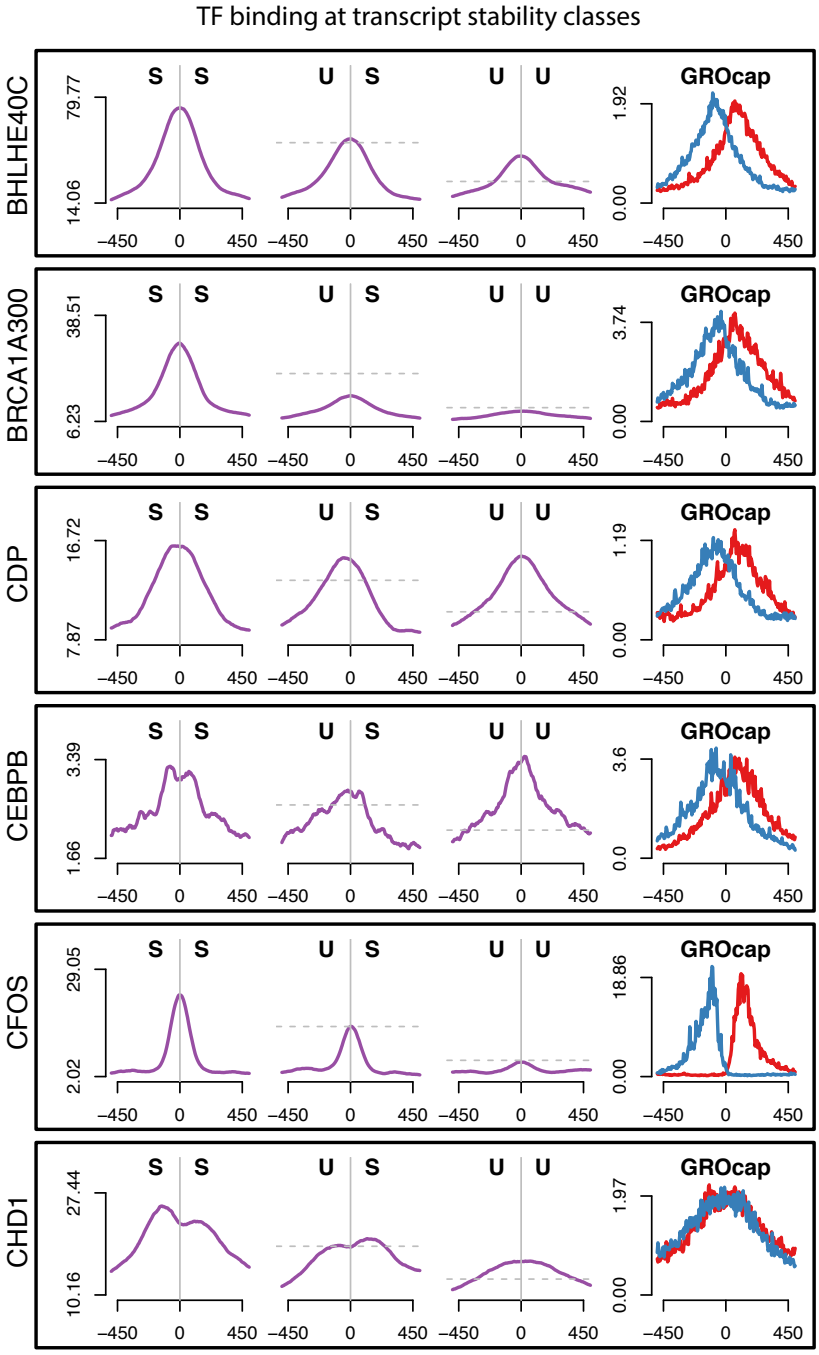

Figure B.14: Profiles of transcription factors at TSS pairs after stability classification (2/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

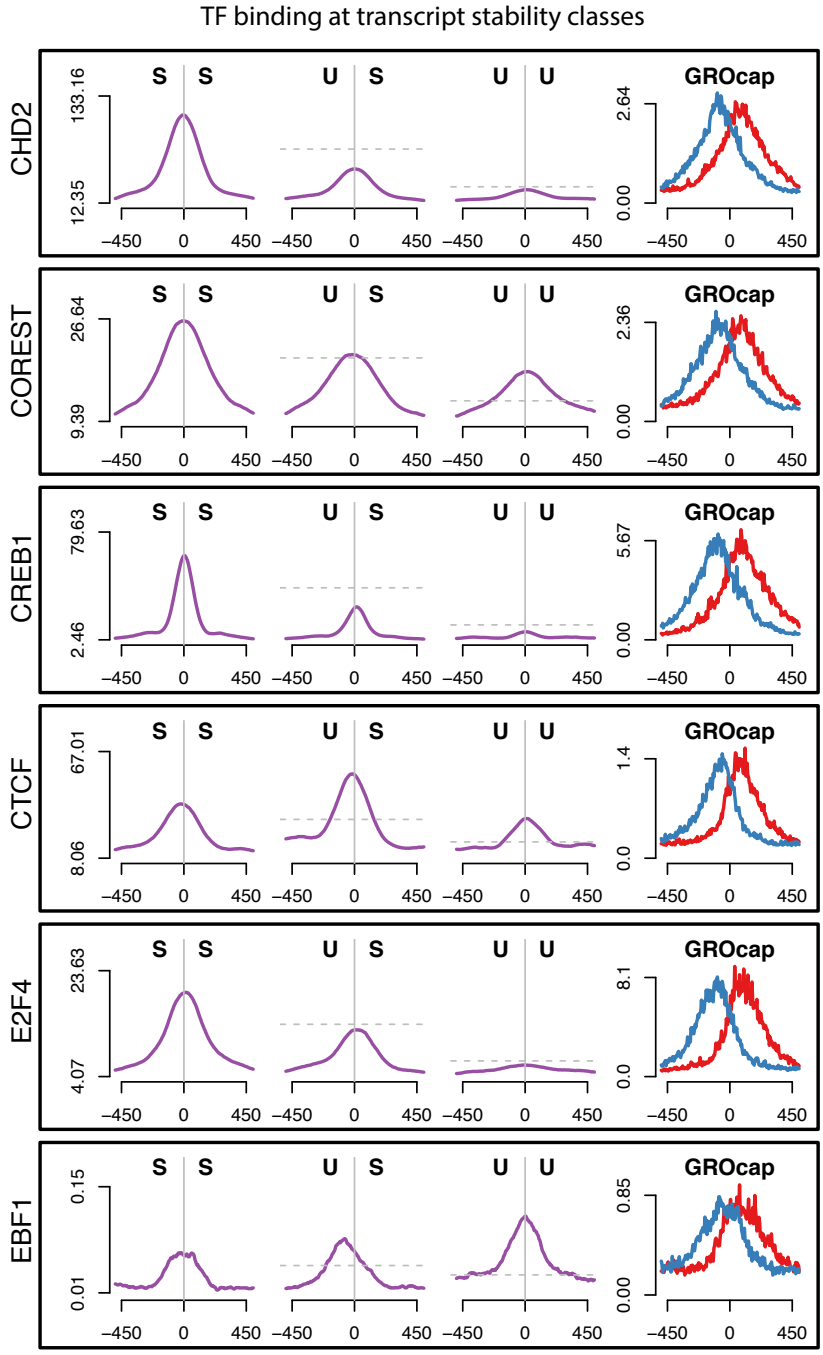

Figure B.15: Profiles of transcription factors at TSS pairs after stability classification (3/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

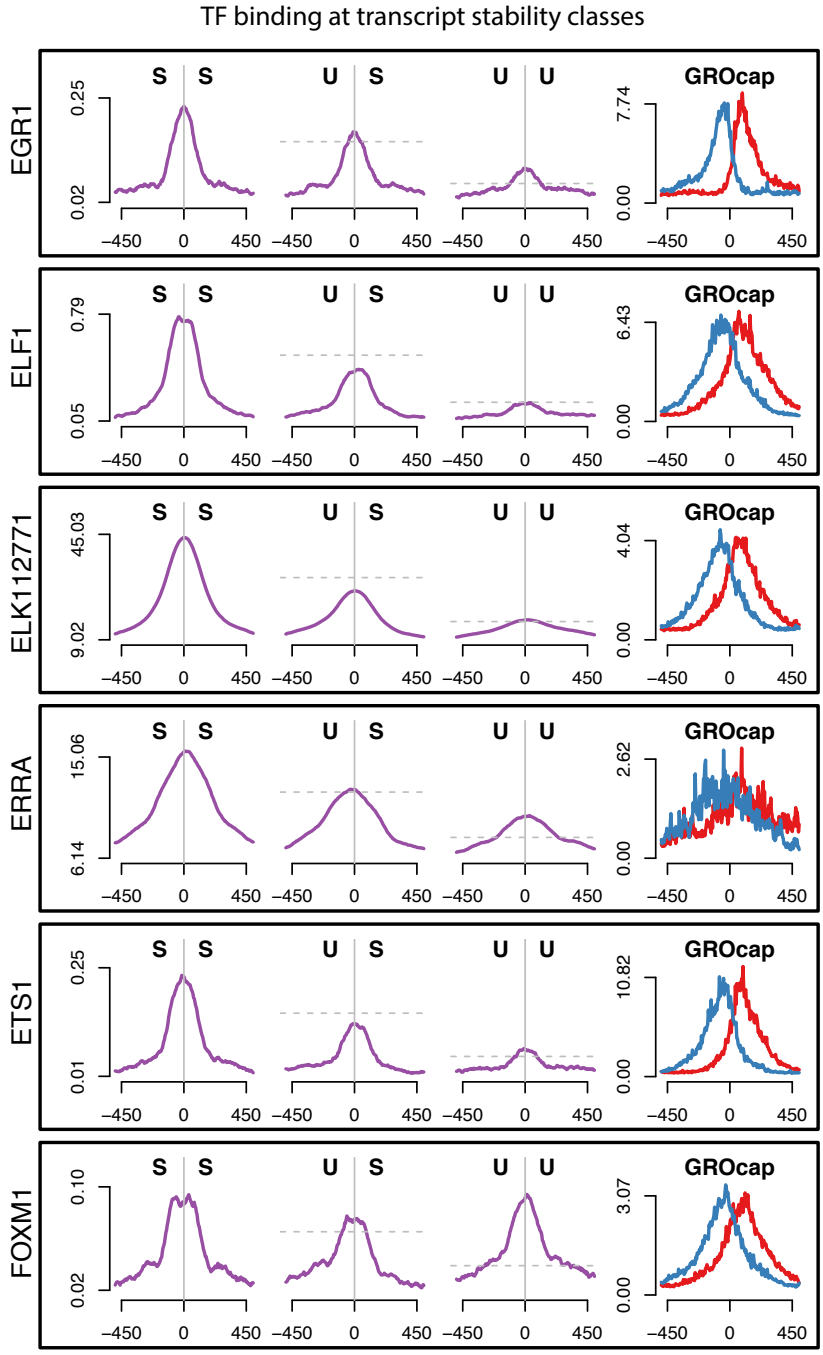

Figure B.16: Profiles of transcription factors at TSS pairs after stability classification (4/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

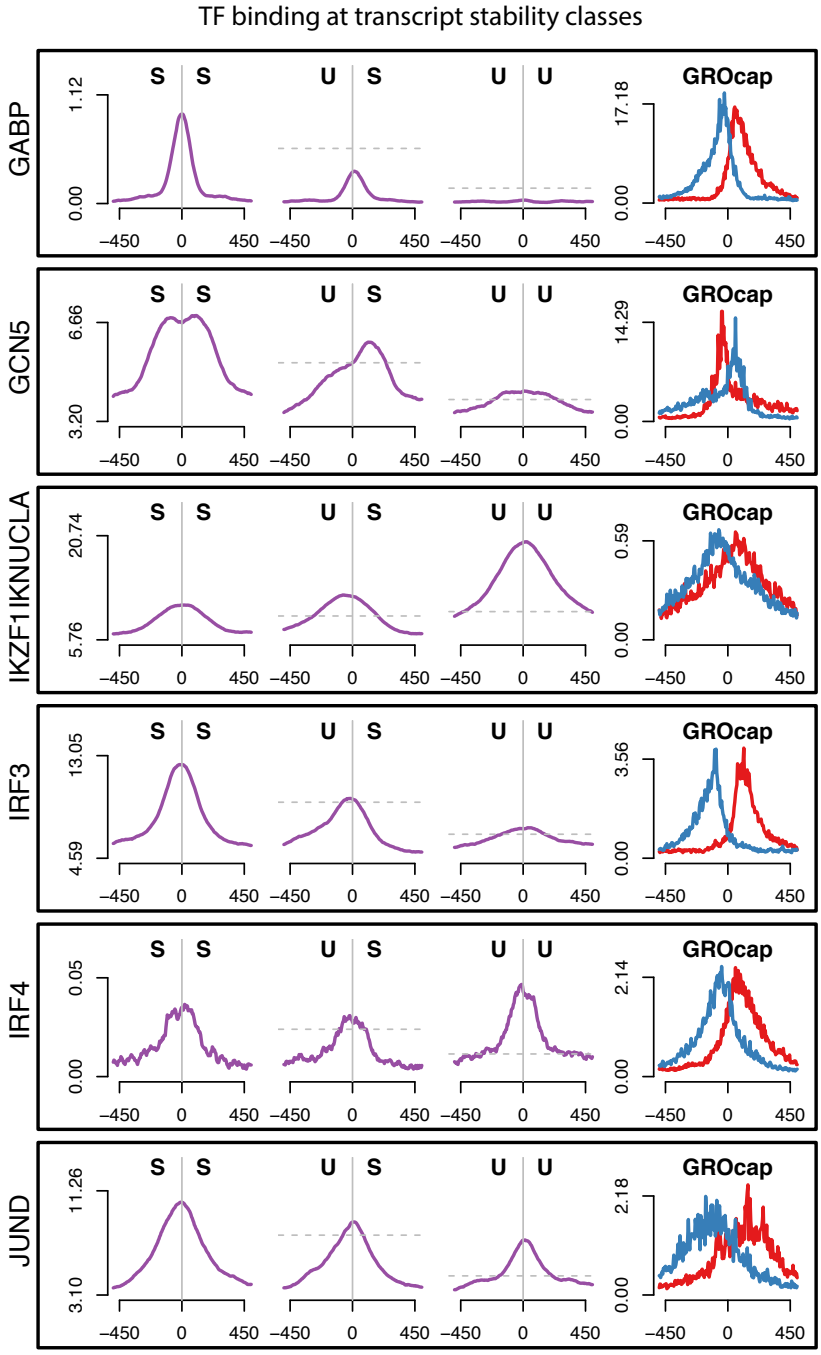

Figure B.17: Profiles of transcription factors at TSS pairs after stability classification (5/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

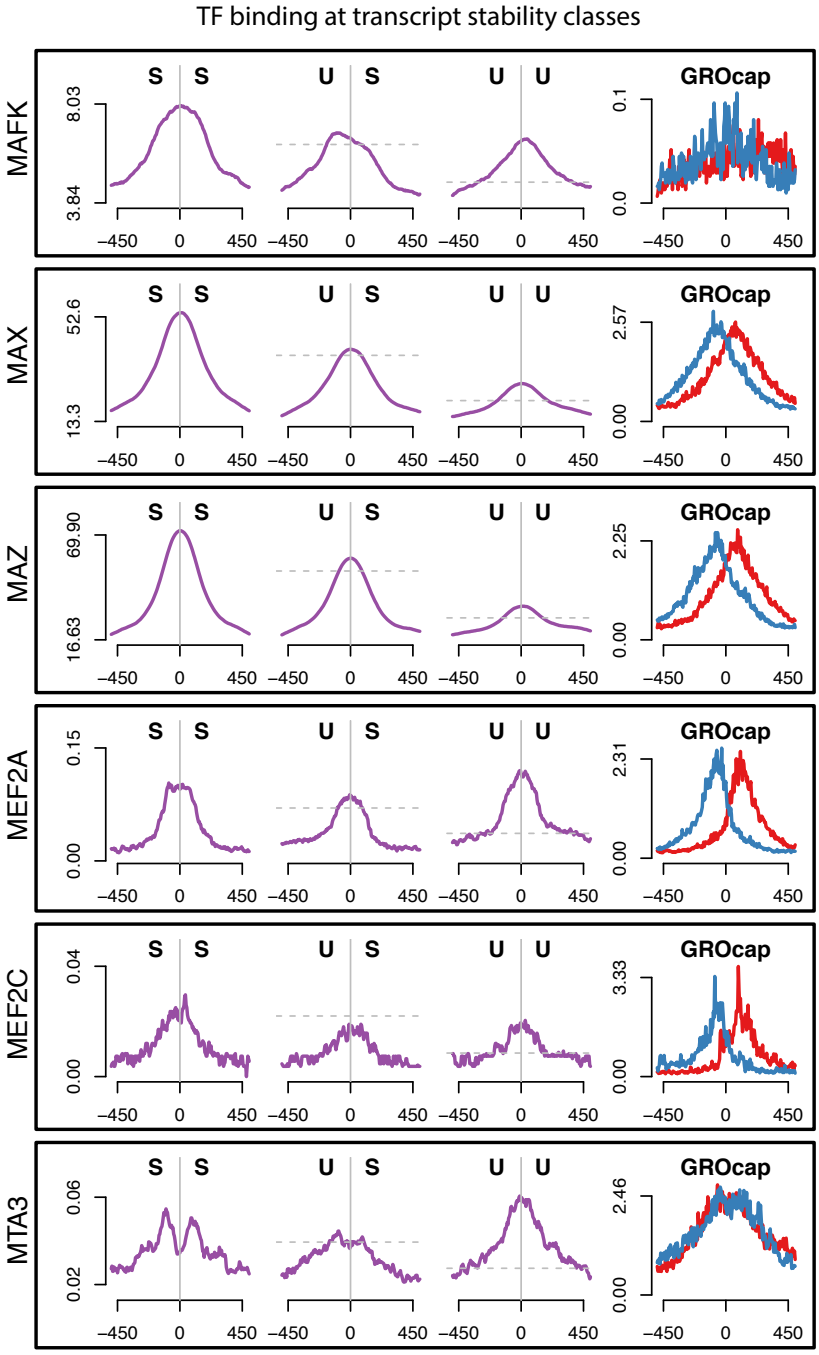

Figure B.18: Profiles of transcription factors at TSS pairs after stability classification (6/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

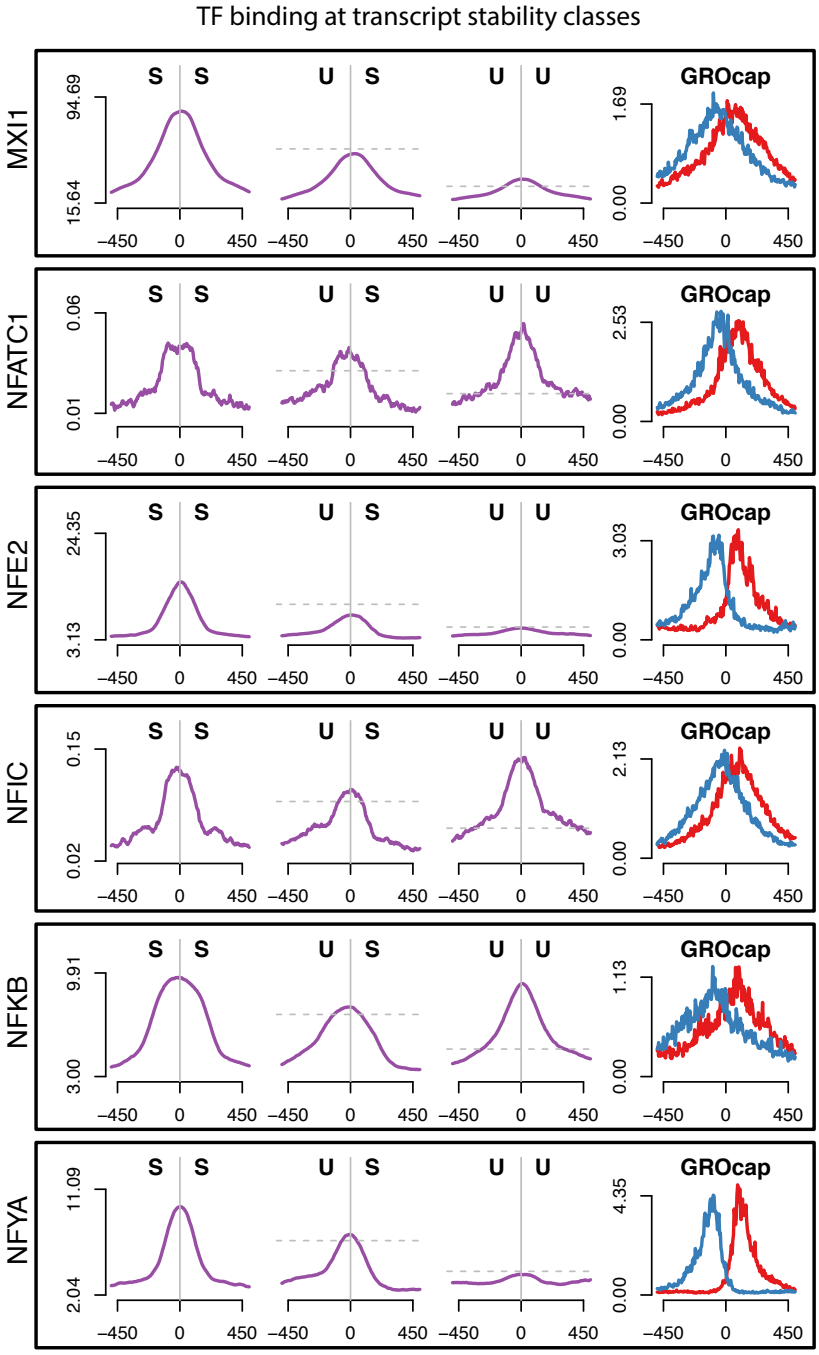

Figure B.19: Profiles of transcription factors at TSS pairs after stability classification (7/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

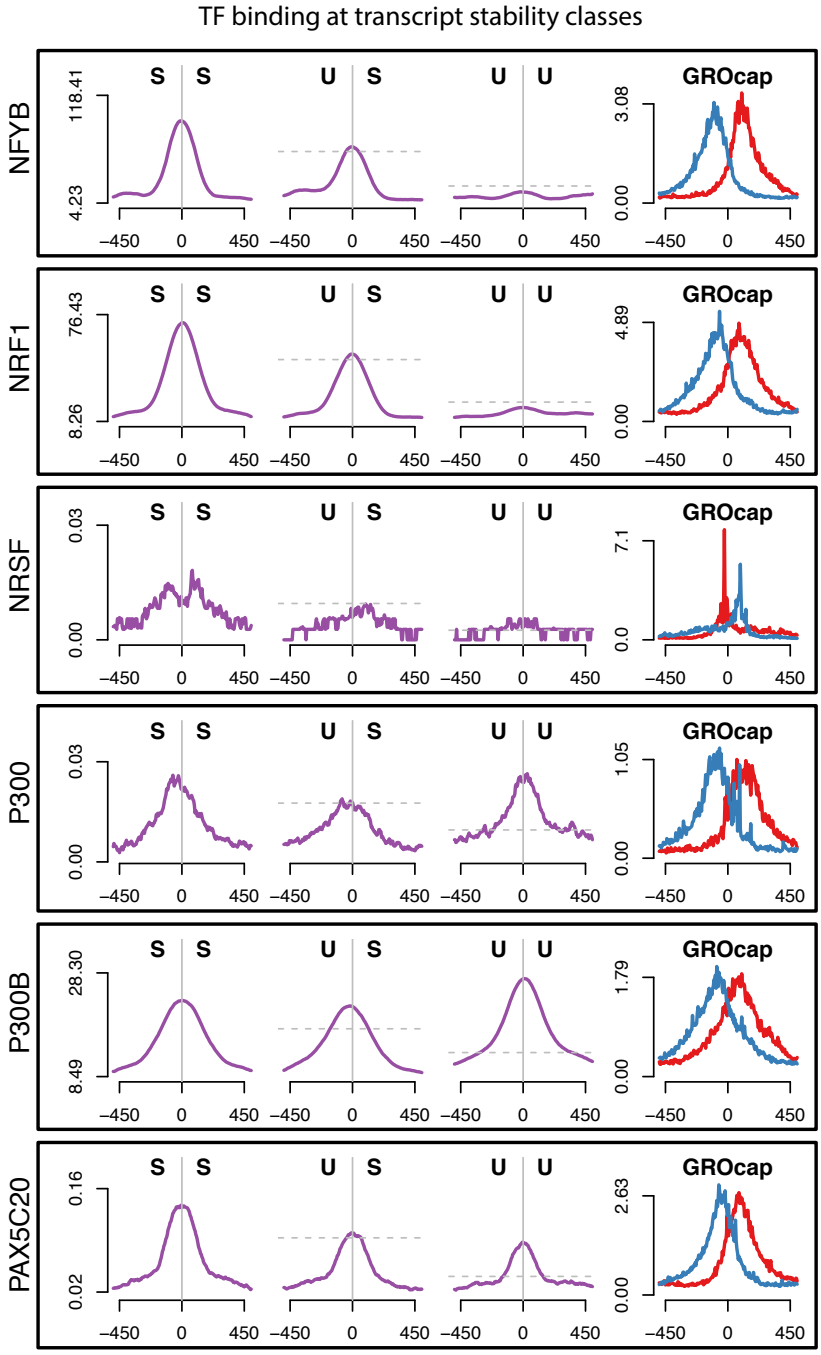

Figure B.20: Profiles of transcription factors at TSS pairs after stability classification (8/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

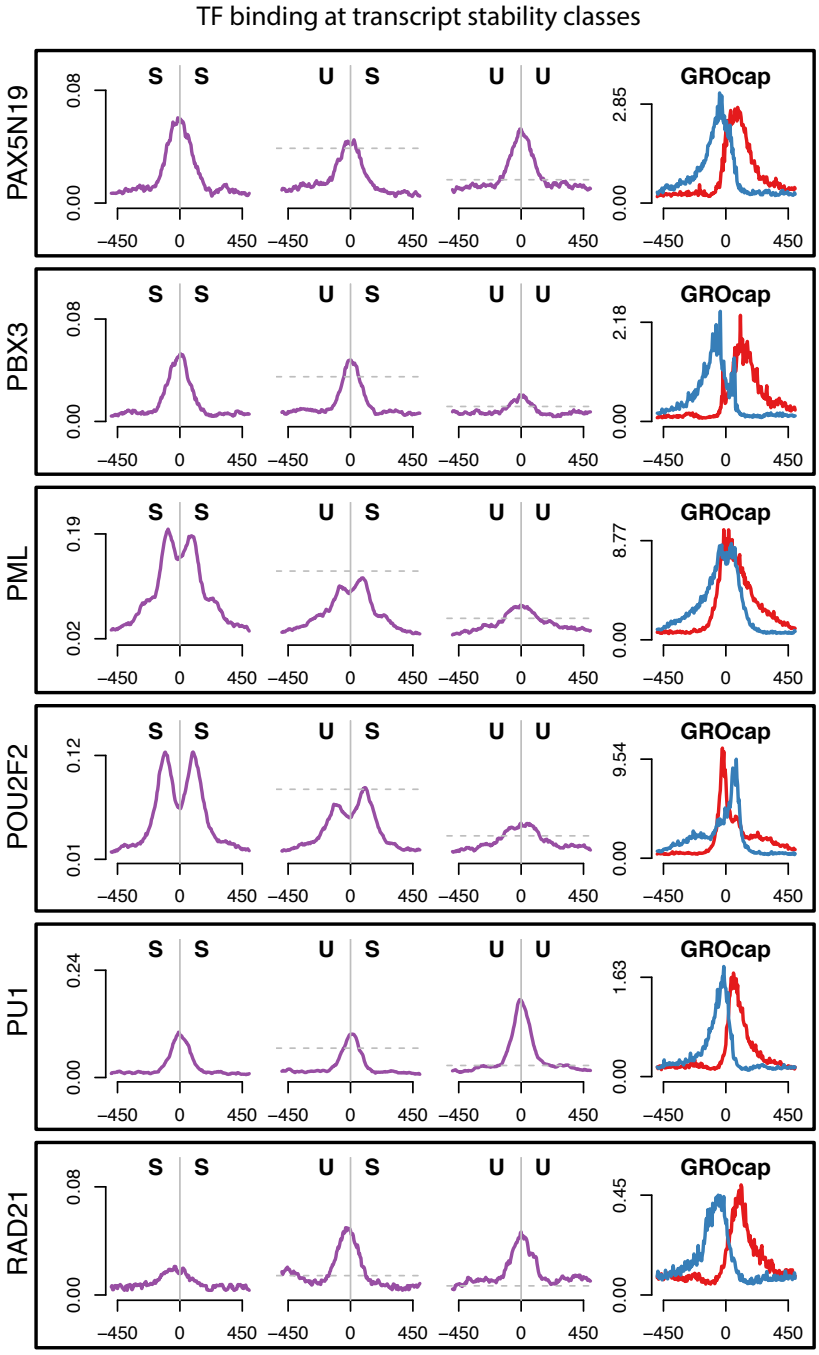

Figure B.21: Profiles of transcription factors at TSS pairs after stability classification (9/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

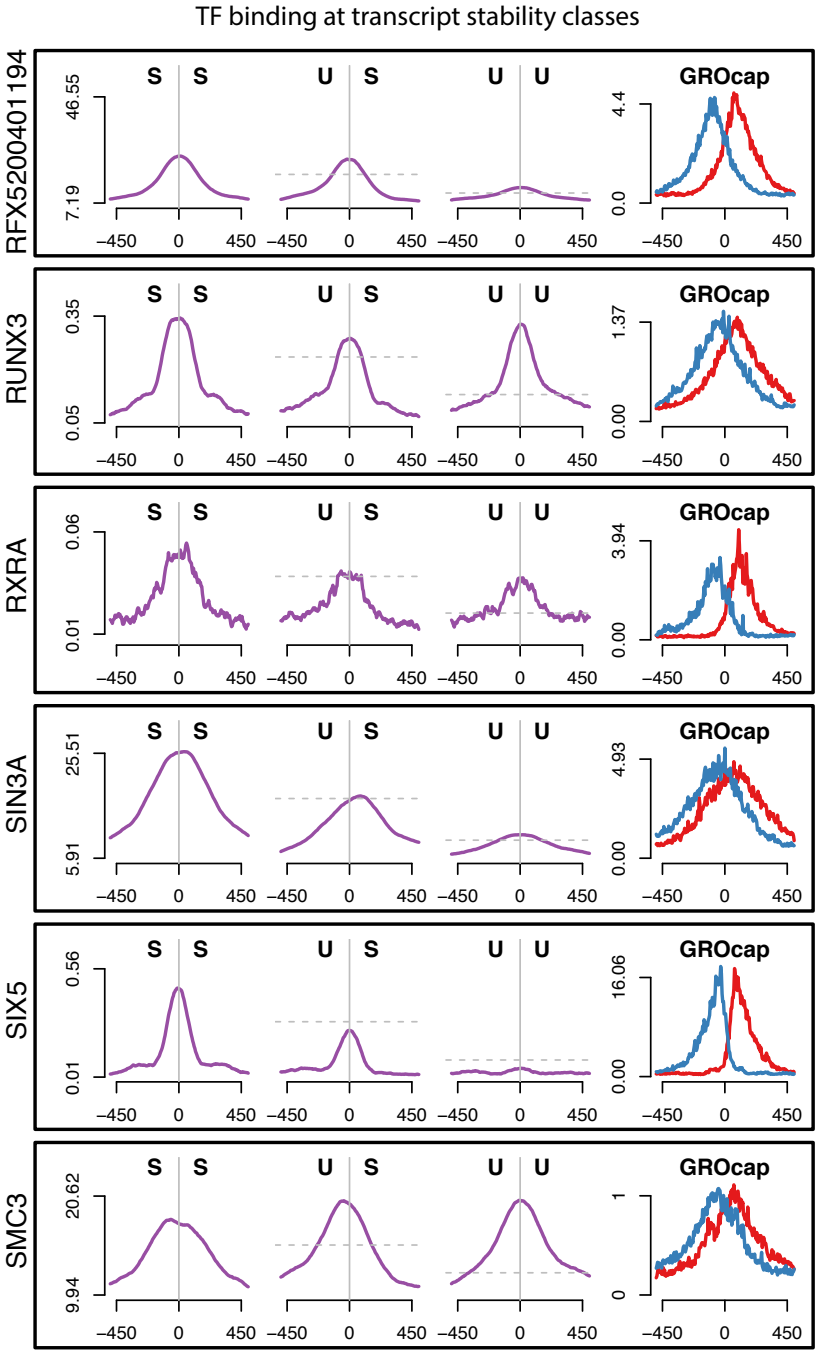

Figure B.22: Profiles of transcription factors at TSS pairs after stability classification (10/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

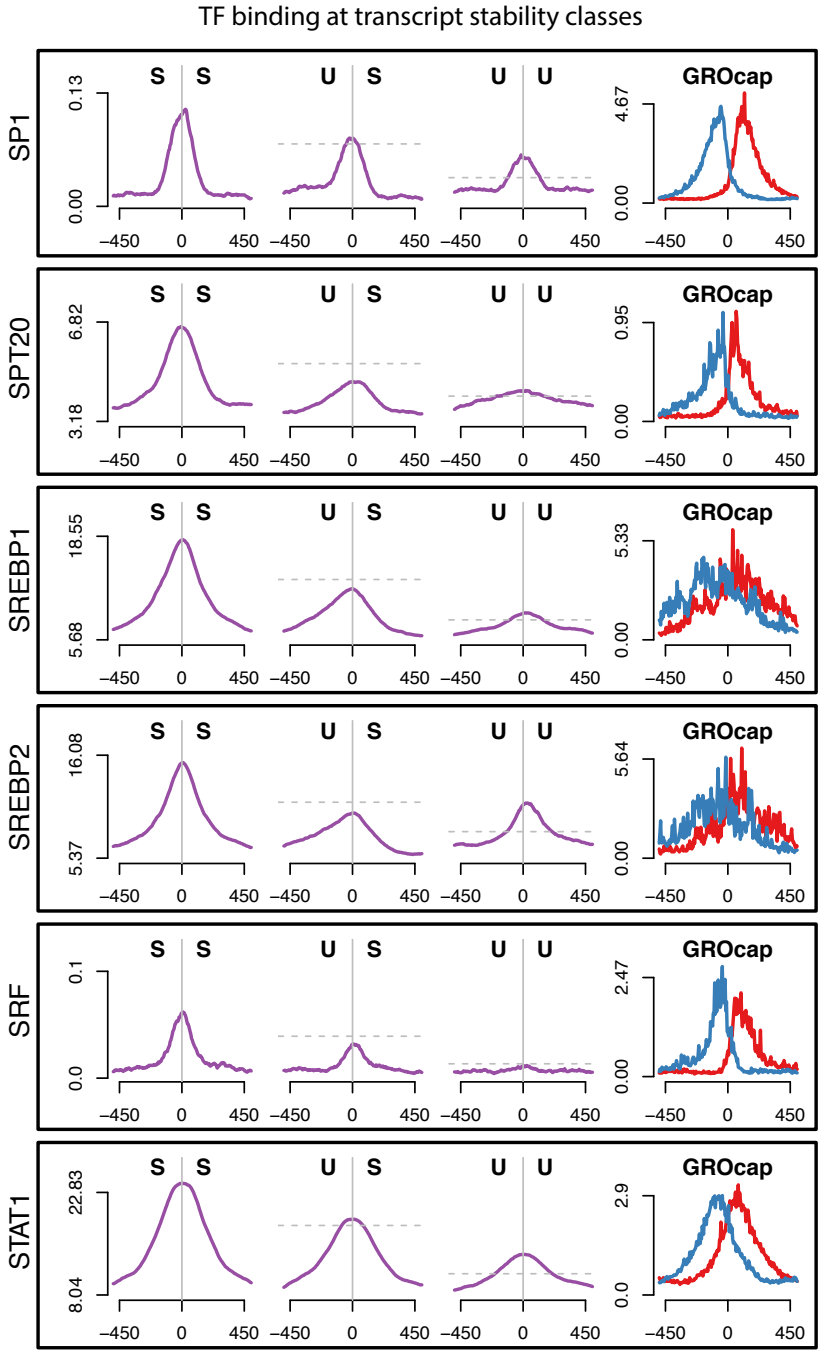

Figure B.23: Profiles of transcription factors at TSS pairs after stability classification (11/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

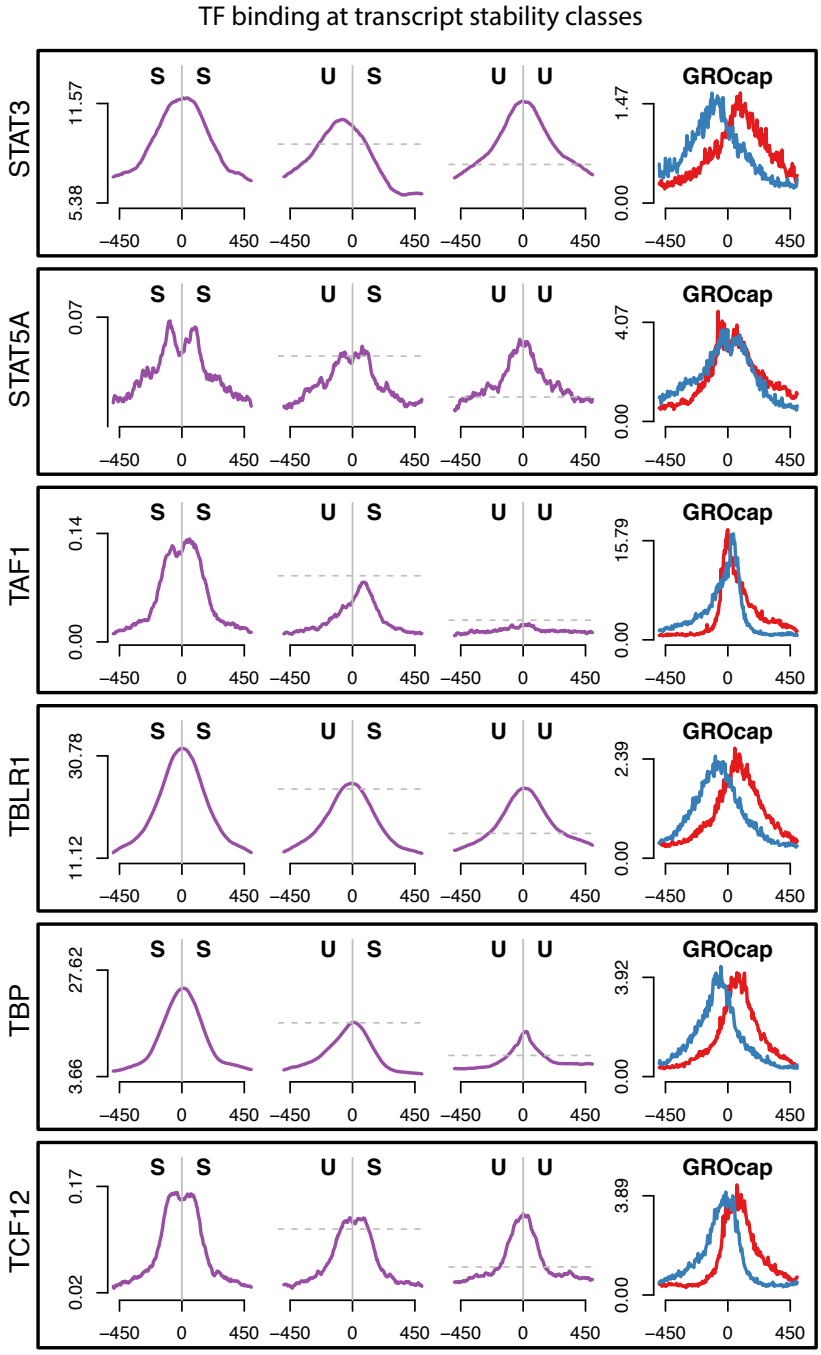

Figure B.24: Profiles of transcription factors at TSS pairs after stability classification (12/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

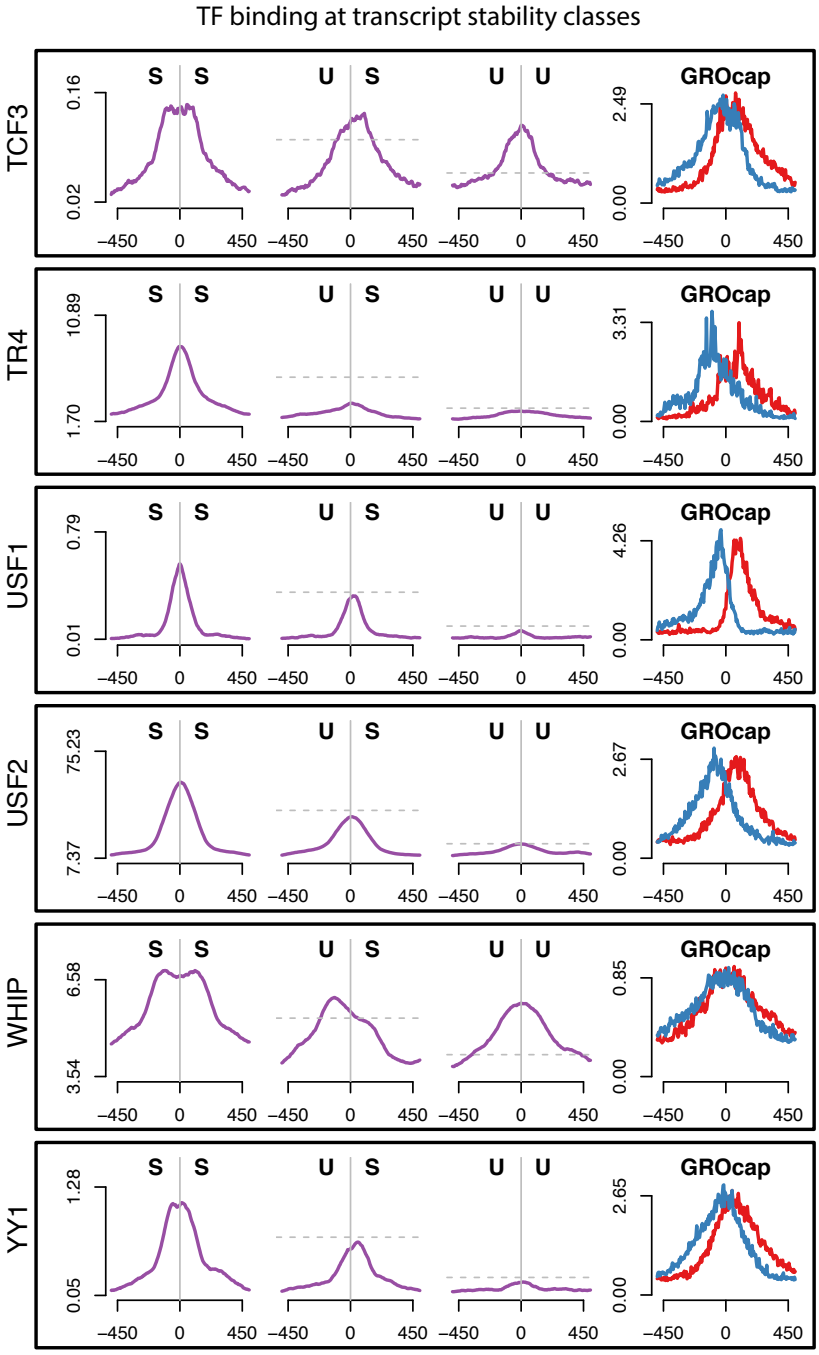

Figure B.25: Profiles of transcription factors at TSS pairs after stability classification (13/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

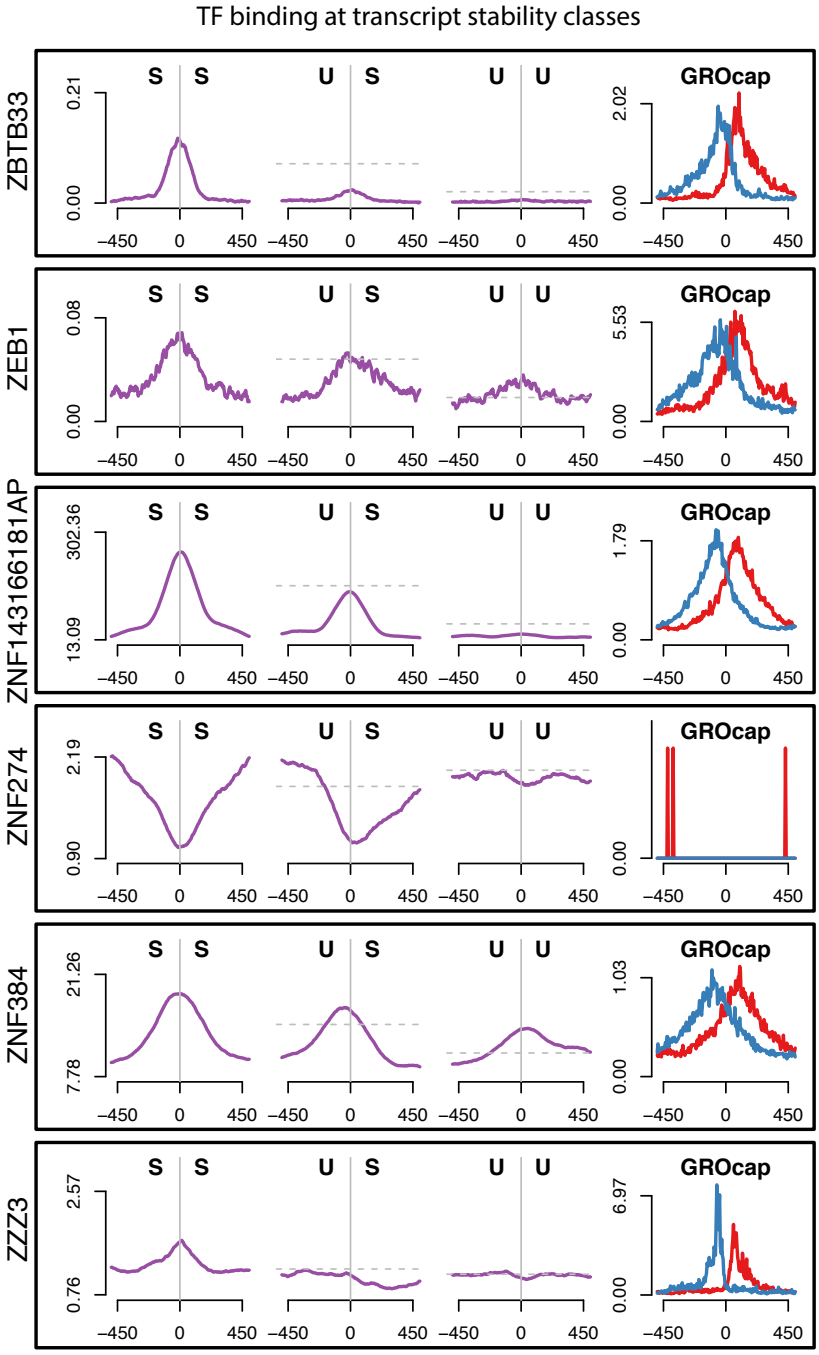

Figure B.26: Profiles of transcription factors at TSS pairs after stability classification (14/14). Composite profiles of ChIP-seq data for various transcription factors aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable, unstable::stable, unstable::unstable. The horizontal dashed lines represent the expected peak signal level if the signal followed the scaling of Pol II relative to the SS panel. The right panel shows GRO-cap data aligned to the peak of each individual transcription factor. All ChIP-se data was produced by the ENCODE consortium in GM12878 cells.

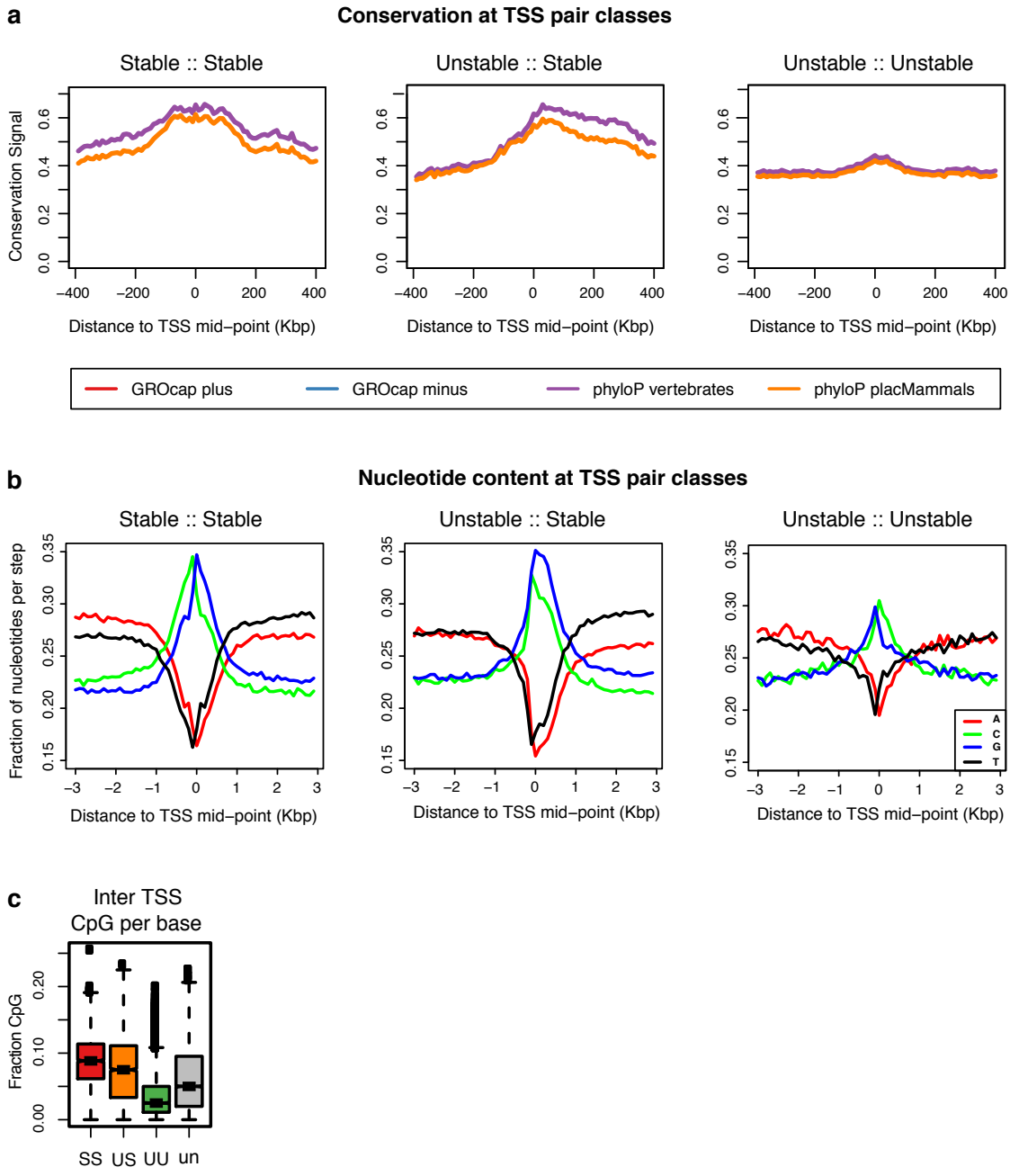

Figure B.27: Sequence conservation and composition. (a) PhyloP [108] score for vertebrates (purple) and placental mammals (orange), aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. (b) Nucleotide frequencies aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. (c) Fraction of CpG dinucleotides in between divergent TSSs in pairs for different stability classes.

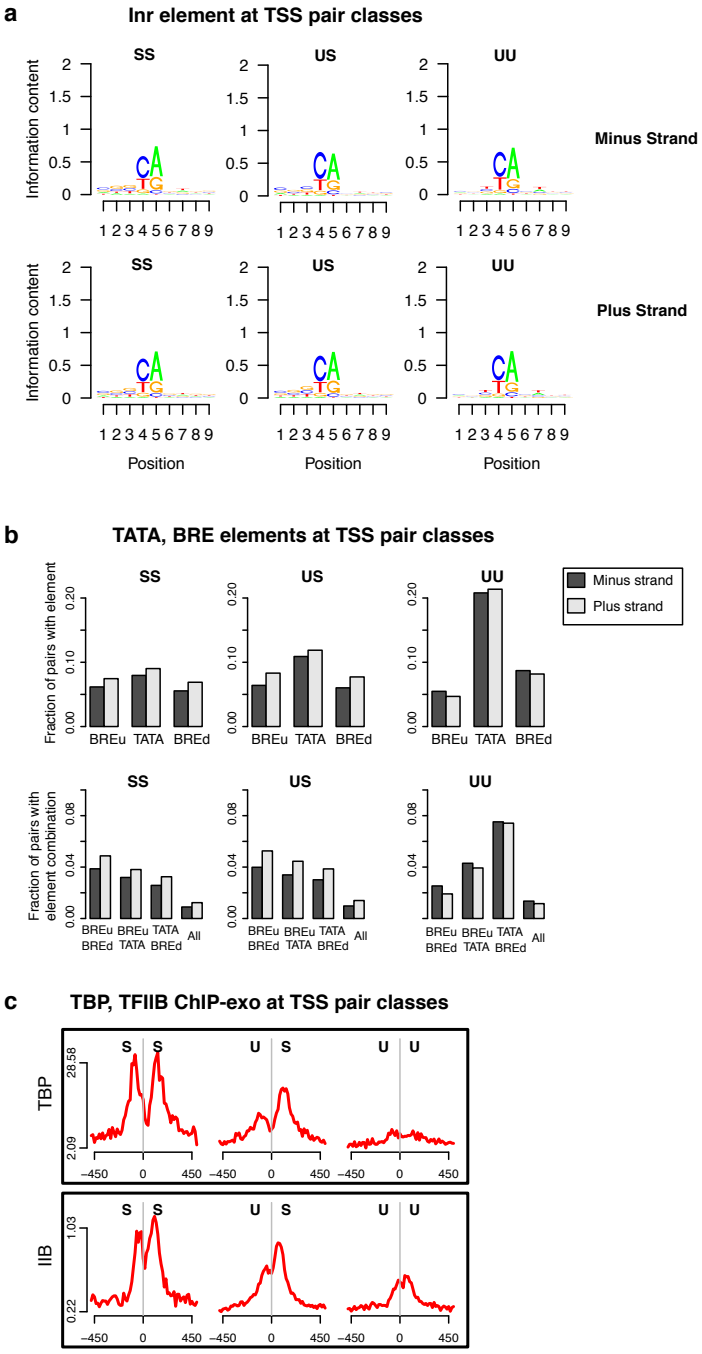

Figure B.28: Sequences at TSS. (a) Sequence logos showing INR element underlying both minus strand (top) and plus strand (bottom) TSSs at the different transcript stability classes. Logos obtained by alignment on base with strongest GRO-cap signal in each TSS region. (b) Occurrences of core promoter elements (TATA, BREd, BREu) at canonical positions [138]. Top shows individual elements and bottom row shows combinations of elements. (c) ChIP-exo profiles for TBP and TFIIB [138] (K562 cells) aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Profiles are stable::stable (left), unstable::stable (center), unstable::unstable (right).

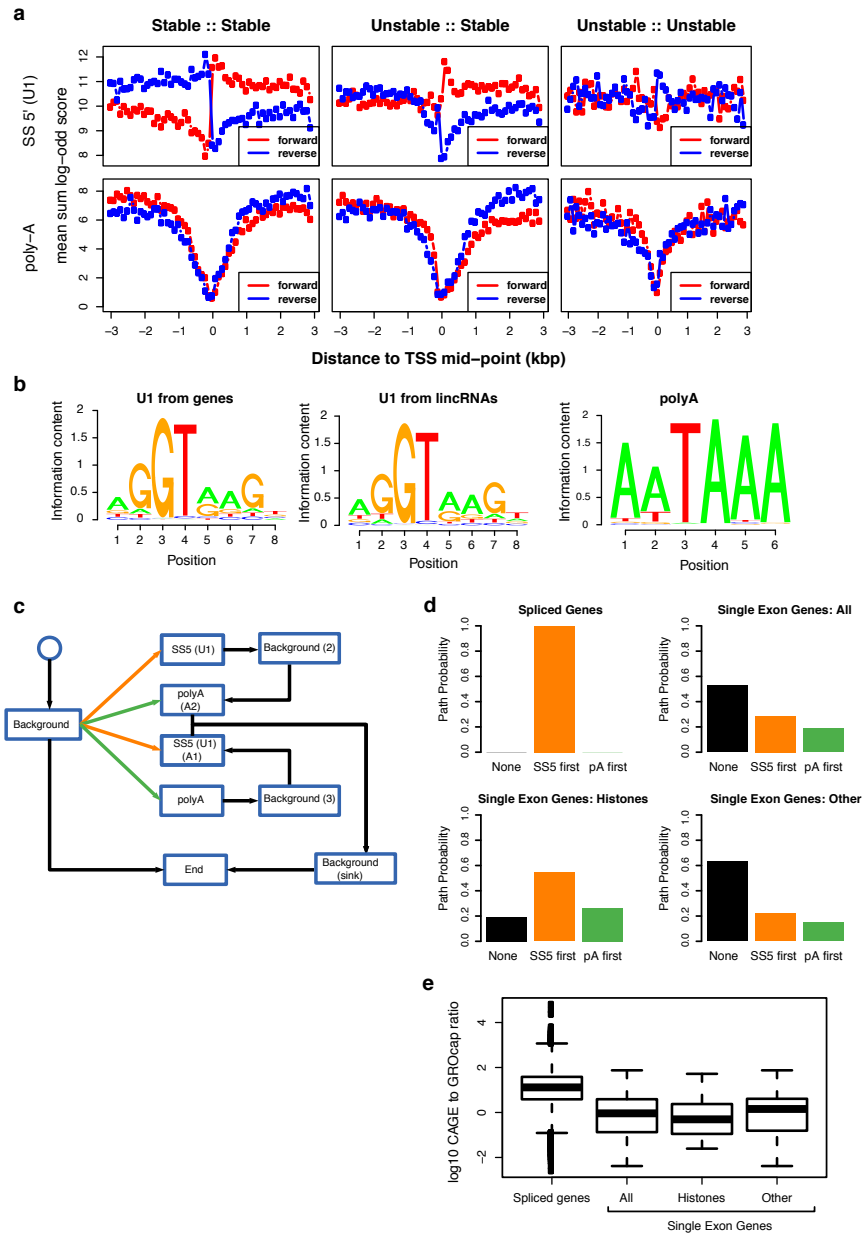

Figure B.29: (a) Five-prime splice site (SS5; top) and poly-A (PAS; bottom) log-odds score profile in forward (red) and reverse (blue) strands aligned to the center of GRO-cap TSS pairs after classifying pairs based on the stability of the transcript produced. Consistent with previous work [3][100], the SS5 motif is enriched downstream of TSSs that produce stable transcripts, but depleted at unstable transcripts. In contrast, the PAS motif is depleted downstream of stable TSSs. (b) PWM motifs for SS5 and PAS elements used in (a). SS5 PWMs obtained from GENCODE annotations with no apparent difference between protein-coding and lincRNAs. PAS PWM from [12]. (c) HMM diagram for PAS versus SS5 relative motif position analysis. Boxes represent sequences of states representing the corresponding PWM motifs. Alternative paths capture the various possible relative element positions. (d) Estimated path posteriors through HMM for spliced gene transcripts and curated single exon gene transcripts. Single exon set is further split between histone coding transcripts and other. (2) GRO-cap to CAGE ratios in the subsets shown in (d).

Table B.1: Summary of datasets and mapped reads generated for this study

Cell line	Assay	TAP used?	Length of mapped reads (bp)	# reads mapped
GM12878	GRO-cap	no TAP	30	6541296
GM12878	GRO-cap	with TAP	30	27314798
GM12878	GRO-seq	N/A	30	105765321
K562	GRO-cap	no TAP	30	9267605
K562	GRO-cap	with TAP	30	26634162
K562	GRO-seq	N/A	30	12721755
K562	PRO-seq	N/A	15-100	364790421

Table B.2: Classifications from the literature associated with TFs found in the 'TSS cluster'

Factor	Repressor	Activator/Co-act.	GTF
Chd1			
Gcn5		×	
Mta3	×		
Nrsf	×		
Pml	×		
Pou2f2	×		
Stat5a	×		
Taf1			×
Whip			
YY1	×	×	

APPENDIX C

SUPPLEMENTAL MATERIAL FOR CHAPTER 4

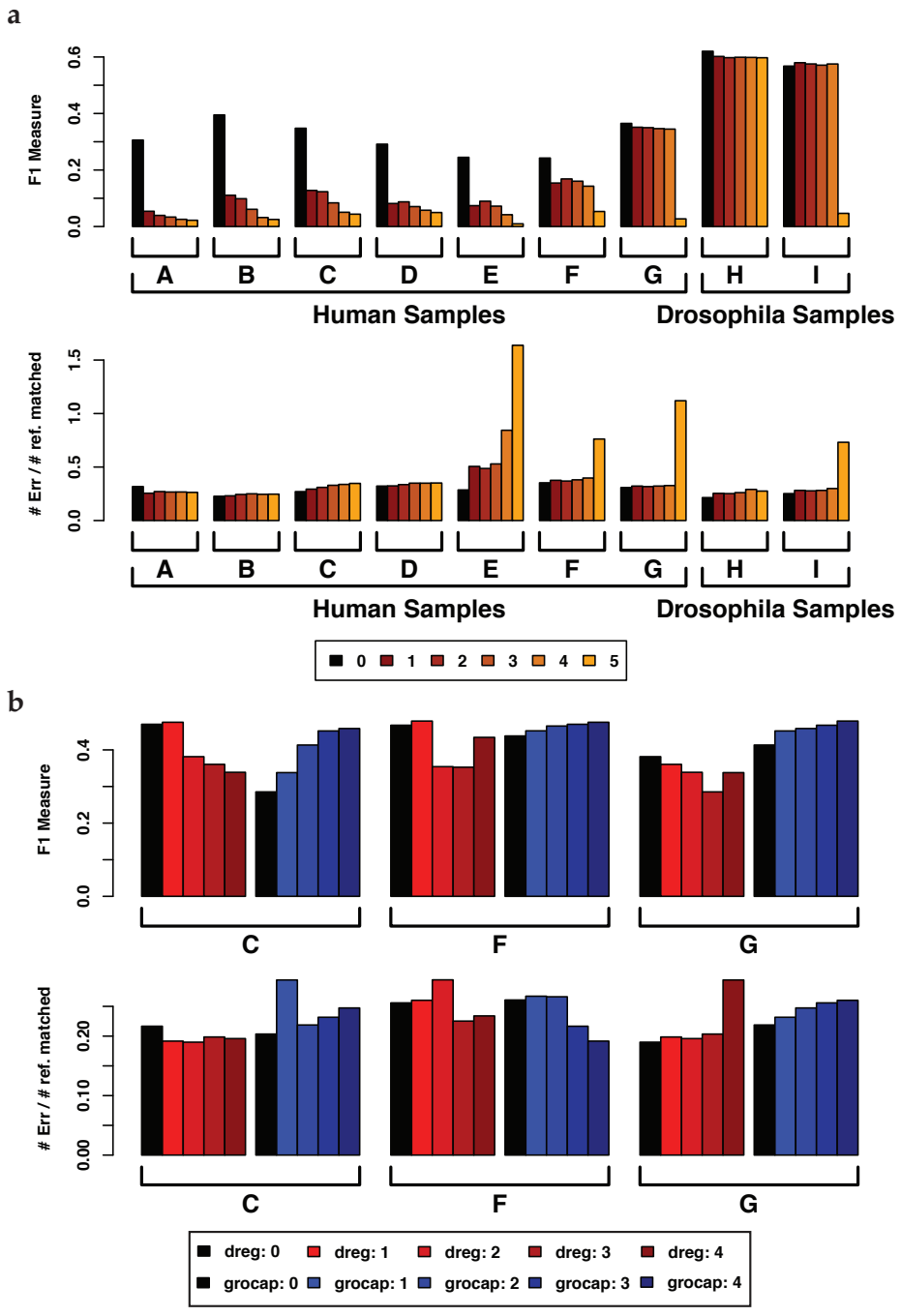

Figure C.1: TU HMMs with multiple transcript paths. (a) (1+2K)-state HMM without any TSS with signal; K varies from 0 to 5. (b) (1+2K)-state HMM with either dREG or GRO-cap TSS region signal ($\gamma = 0.1$); K varies from 0 to 4.

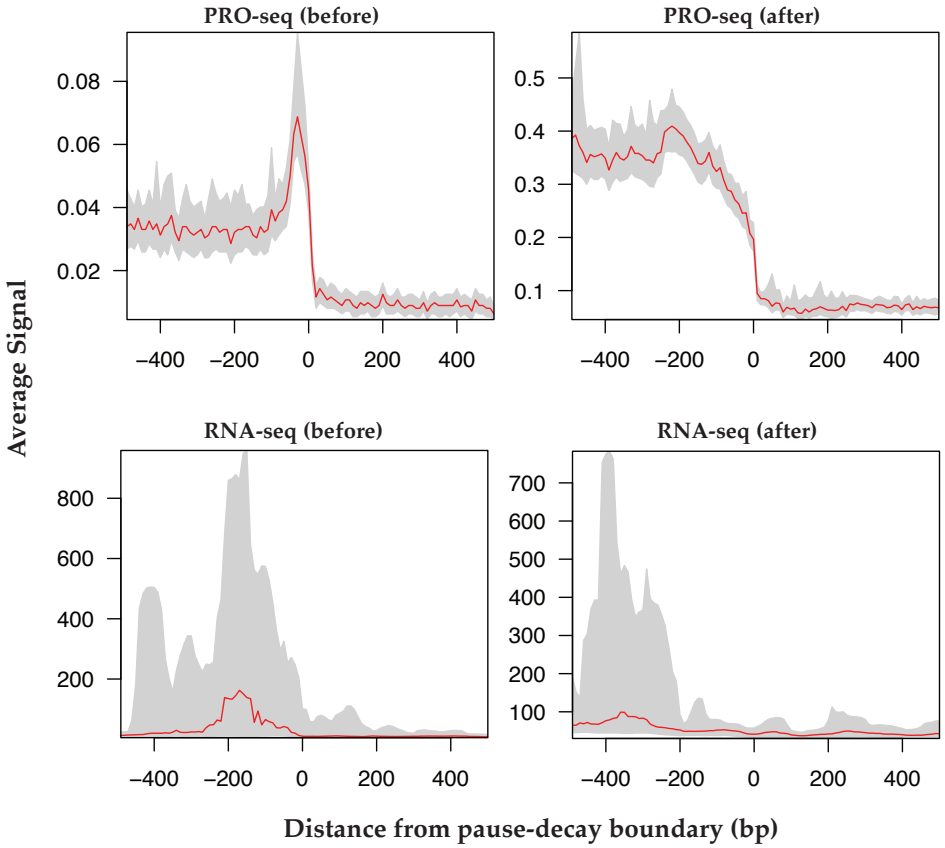

Figure C.2: Profiles at pause-decay boundary. Top row shows PRO-seq profiles before (left) and after (right) the pause edge adjustment. Bottom row shows Nucleus RNA-seq profiles before (left) and after (right).

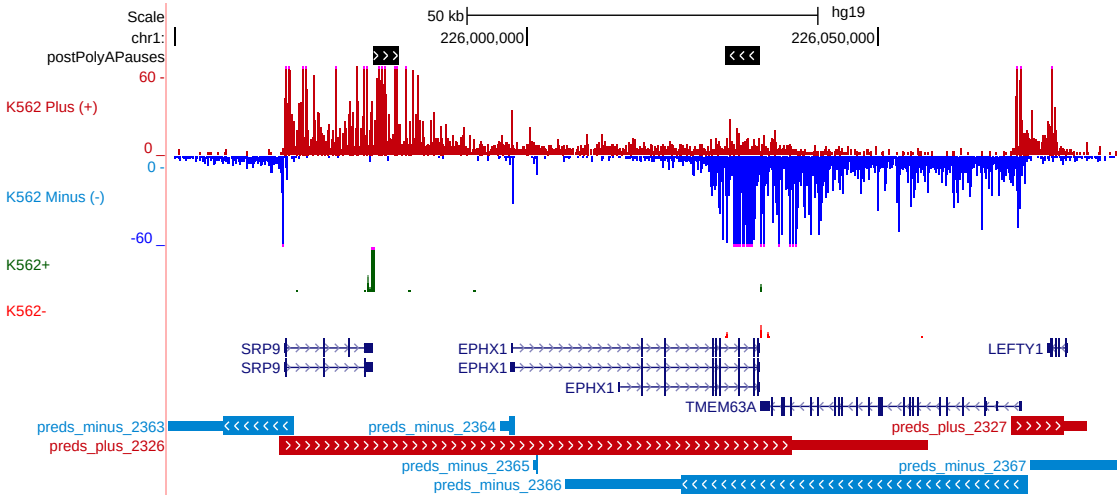

Figure C.3: Post-polyA pause area example. Refined post-polyA pause regions indicated by black boxes in the top track.

Table C.1: Size of TU reference sets. Long references take data from ENCODE/GENCODE (human) and modENCODE (Drosophila) sources for annotations, RNA-seq and CAGE. Human short references are the result of intersecting DNAse HS peaks (OpenChromatin Consortion) and H3k27ac peaks (ENCODE Broad set for the first three, ENCODE SYDH set for MCF-7, Epigenome Atlas for IMR90; humans only). Drosophila short references use only DNAse HS peaks (SRA SRP010823).

Cell Line	Long TU	Long TU groups	Short TU
hg19/gm12878	11566	9108	44716
hg19/k562	11274	8678	43926
hg19/hela	10138	8044	44083
hg19/imr90	14010	10461	30016
hg19/mcf7	13887	10460	14870
dm3/s2	1739	1721	9212

APPENDIX D

BIGWIG R PACKAGE

D.1 Introduction

The rapid development of next-generation sequencing technology, and derived assay methods (e.g. ChIP-seq), lead to the rapid growth of genomic dataset file sizes, with typically over a gigabyte per file. The need to aggregate, visualize and query such large datasets has spurred the development of web data hubs like the UCSC Genome Browser [75] and, later, the creation of efficient file formats like the bigWig file format [76]. BigWig files are compressed binary indexed files containing data at several resolutions to allow efficient query by the UCSC Genome Browser.

The UCSC Genome Browser is accompanied by a set of command line tools that enable rudimentary manipulation of the file formats it uses, but fundamentally they are geared towards browser usage. The need to access bigWig files, now the common format for processed genomic datasets, has spurred the creation of both independent data analysis programs (WiggleTools [148]), command line tools (bwtools [107]), as well as packages for some common scripting languages (Bio-BigFile for Perl [129], bx-python for Python [133]).

The bigWig R package was created by the present author to enable efficient *read* access to bigWig files from within the R programming language, enabling the easy integration of genomic data with the statistical analysis packages available to R. It is built on top of the UCSC Genome Browser bigWig C libraries and features an extended set of query functions and associated utilities to facilitate

large scale data access. Furthermore, the core of the bigWig R package offers an intermediate level C API (Application Programming Interface) that can be used to write other software packages that require access to bigWig files at a higher level than that which is provided by the UCSC Genome Browser code.

D.2 Description

The bigWig R package defines two types: *bigWig* and *bwMap*. The latter encapsulates a bigWig file containing per position mappability information for a given read size (unmappable positions encoded as 1 and mappable positions can either be encoded as zeros or omitted from the bigWig file) for use as additional information in query functions.

At the core, there are three functions for each type ([.] denotes an optional name component, (.|.) denotes alternative name components, and (...) denotes the omitted function arguments):

```
(load|unload|print).bigWig(...)
(load|unload|print).bwMap(...)
```

which enable loading, unloading and displaying of objects of the corresponding types. The `print.bigWig` function will display the same information as presented by the UCSC Genome Browser command line tool `bigWigInfo`. The `load.(bigWig|bwMap)` functions support loading bigWig files from HTTP URLs efficiently, with only the queried portions being actually downloaded (remote querying is one of the core features of the bigWig design [76]). In cases where bigWig datasets are split among multiple files (one file per chromosome),

the query functions described below will also accept a path prefix/suffix pair (as a character vector; file path defined as `<prefix><chrom><suffix>`) in place of bigWig objects. Given instances of the core objects, the package supplies a few sets of useful functions:

Query functions There is a simple query function that provides access to the underlying data representation (`query.bigWig`), as well as a set of functions that provide higher level access. Higher level functions are named by joining a set of name components to determine the desired semantics:

```
[(bed|bed6).](region|step).(bpQuery|probeQuery).bigWig(...)
[(bed|bed6).](region|step).bpQuery.bwMap(...)
```

From right to left, after the object type, the first component determines how value intervals[1] are interpreted: `bpQuery` interprets the data from a single base pair perspective (even if data is stored using intervals) and `probeQuery` interprets each interval as a single value. Note that `bwMap` only exposes `bpQuery` functions subset. The next name component determines if a query range is to be reported as a single value (`region`) or if it should be split into steps of identical size (`step`). Finally, the optional leftmost name component, determines that instead of a single query range, the function should be repeated for all query ranges present in an R data frame that has the contents of a BED file. Use `bed` if the query ranges have no strand information and `bed6` if there is strand information (when using strand information, two bigWig objects must be passed to the function, one for the plus strand and one for the minus strand). All these functions

[1] At its core, bigWig files correspond to a set of non-overlapping coordinate intervals, each with one associated value

then support query operators such as min, max, sum and average, that determine how values are aggregated in each step or region.

BED functions BED data, read into R as a data frame, is the most common way to specify query ranges. As such, the package provides a series of utility functions to perform simple transformations on these data frames (`center.bed`, `fiveprime.bed`, `downstream.bed`, `upstream.bed`, `threeprime.bed`). It also supplies a function that applies a user supplied function to each row in a BED data frame (`foreach.bed`).

Profile functions A common operation with genomic data is to create signal profiles that aggregate data across many similar locations across the genome. The bigWig package provides various utility functions to help generate these profiles, as well as plot them. Profiles can be generated that correspond to signal quantiles (with or without subsampling) or to mean/standard error confidence intervals (with or without bootstrapping).

Altogether, the functions defined in this package, not only facilitate common operations (e.g. creating profile plots that aggregate data from multiple locations), but also serve as a useful building block to write custom functions tailored to the needs of specific genomic data analysis projects.

D.3 Availability

The package source code is available from the present author upon email request (`alm253@cornell.edu`). Compilation requires the UCSC Genome

Browser source, specifically the source for the *jkweb* library, which is available for non-commercial uses from Jim Kent [75].

APPENDIX E

QUICK HMM R/C++ PACKAGE

E.1 Introduction

Hidden Markov models (HMMs) are a commonly used statistical framework to model data sequences. There are various packages available to define and apply these models, each with their own set of constraints (see [91] for a recent summary). The library and R package described here aims to provide rapid prototyping, through the R package interface, layered on top of a C++ library that provides an efficient implementation of the core algorithms, as well as extension points to add new emission or transition function distributions. This significantly extends the previous available HMM package in R [64] while providing a transition path between an initial R based prototype and a full independent program.

The Quick HMM (QHMM) code supports discrete-time first-order HMMs, with both homogenous and non-homogenous transitions. Emissions can be uni- or multi-dimensional, with multiple conditionally independent emission tracks. Furthermore, emission tracks can have missing data (specified separately) and, both emissions and transition distributions can make use of extra (fixed) covariate data. Finally, parameters can be selectively fixed during EM and parameter sharing can be achieved by specifying groups of states, for transitions, or groups of state/track pairs, for emissions, that share a common distribution class (but independent class instances). The code provides implementations of the following algorithms: expectation maximization (EM) (core code, each transition/emission distribution must supply its corresponding part of the

EM loop), forward, backward, viterbi, state posteriors, transition posteriors and stochastic backtrace.

The current library has implementations for several emission distributions (discrete, discrete gamma, geometric, poisson, negative binomial) and transition distributions (discrete, logistic regression, auto-correlation and mixtures of auto-correlation with covariate based priors).

E.2 C++ Library Architecture

The C++ library is designed with the goal of providing an efficient implementation of the core algorithms described above, while maintaining a simple and flexible interface for both the user of the library and the developer of new emission and transition distributions.

The balance between performance, flexibility and ease of development of new distributions rests on three aspects: leveraging C++ templates to both provide alternative implementations for core components and to shift polymorphism costs from runtime to compile time; structuring code to avoid paying the extra runtime cost for unused features; and taking advantage of parallelism by the use of threads (implemented via OpenMP).

In particular, C++ templates are used to support polymorphism in implementation choices such as: homogenous versus non-homogeneous transition tables; single versus multi-track emissions; dense versus sparse transition tables; two-state special cases for log-sum. This enables the compiler to produce efficient code targeted at the specific choice of features in use.

Some features require additional runtime checks, instead of alternative implementations, as such the code is organized to avoid those costs when the features are not active, either by use of object wrappers (for missing data and debug checks), or via local caches (transition table is cached in the homogenous case, avoiding the function call cost required to implement non-homogeneous HMMs).

Finally, care is taken to isolate this extra complexity from the implementation of emission and transition distributions by isolating parallelism in the core code, as well as using iterator objects to keep details such as missing data in emissions and memory management aspects of posterior computation hidden from distribution developers.

E.3 R Package Description

The R package provides access to the functionality provided by the C++ library under an easy to use interface. Usage follows a three step process: (1) creating the HMM object, (2) setting the initial parameters and options[1], and (3) invoking the required algorithms (passing in the data as R matrices).

Using the dishonest casino example from Durbin et. al [34] (see Figure E.1) to illustrate these steps, we first define the HMM object. To do so, one must provide at a minimum four pieces of information:

1. the data shape, i.e., the dimension of each emission and covariate track;

[1]The QHMM library distinguishes between distribution parameters subject to inference (simply called parameters) and other constants that control implementation behavior (called options). These option constants are not to be confused with potentially inferable distribution parameters that have been marked as fixed.

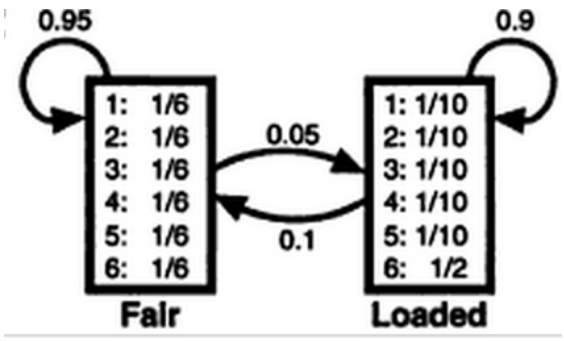

Figure E.1: Dishonest casino two-state HMM from Durbin et. al [34].

in our example there is only one unidimensional emission track (the dice rolls) and no covariates (hence the `NULL` term);

2. the valid transition table - a matrix that defines which HMM state transitions (source is row, destination is column) are valid (marked as positive values) and given them a sequential integer identifier, for each source state. These identifiers are used both by transition distributions to match parameters to transitions and enable matching of transitions within groups for parameter sharing; in our example, the first state has a transition to itself (id 1) and to the second state (id 2), and the second state has a transition to the first (id 1) and to itself (id 2).
3. transition distribution names, one per state; in our example they are simple discrete (also called multinomial) distributions;
4. emission distribution names, one set per state, where each set has one per emission track; in our example, there is only one emission track and again discrete distributions are used for both.

In code, this becomes:

```
hmm = new.qhmm(
        # data shape
        list(1, NULL),
        # valid transitions
        rbind(c(1,2),c(1,2)),
        # transition distributions
        c("discrete", "discrete"),
        # emission distributions
        list("discrete", "discrete"))
```

The next step is to set the various HMM parameters, which is done with the following code (function arguments follow the pattern: HMM object, state number, value vector):

```
set.initial.probs.qhmm(hmm, c(1, 0))
set.transition.params.qhmm(hmm, 1, c(0.95, 0.05))
set.transition.params.qhmm(hmm, 2, c(0.1, 0.9))
set.emission.params.qhmm(hmm, 1, rep(1/6, 6))
set.emission.params.qhmm(hmm, 2, c(rep(1/10, 5), 1/2))
```

Finally, the appropriate algorithm is called (here Viterbi):

```
# load code omitted
# 'rolls' is a vector where die are encoded as the numbers
# 1 through 6
path = viterbi.qhmm(hmm, rolls)
# result 'path' is an integer vector of state numbers
```

More complex examples may require the use of distribution specific options, setting emission and/or transition groups for parameter sharing, passing additional data (covariate or missing data tables), among other things, but this simple example illustrates the overall structure of the R interface.

E.4 Availability

The package source is available from the present author upon email request (`alm253@cornell.edu`). The C++ library can be used independently from the R source, but full feature availability requires linking with the RMath library or the GNU Scientific Library. Multithreaded parallelization requires linking to OpenMP.

BIBLIOGRAPHY

[1] Karen Adelman and John T Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics*, 13(10):720–731, October 2012.

[2] Karmel A Allison, Minna U Kaikkonen, Terry Gaasterland, and Christopher K Glass. Vespucci: a system for building annotated databases of nascent transcripts. *Nucleic acids research*, 42(4):2433–2447, December 2013.

[3] Albert E Almada, Xuebing Wu, Andrea J Kriz, Christopher B Burge, and Phillip A Sharp. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature Genetics*, 499(7458):360–365, July 2013.

[4] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jorgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A Maxwell Burroughs, J Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J Mungall, Terrence F Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O Daub, Peter Heutink, David A Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Mueller, Alistair R R Forrest, Piero Carninci, Michael Rehli, Albin Sandelin, and FANTOM Consortium. An atlas of active enhancers across human cell types and tissues. *Nature Genetics*, 507(7493):455–461, March 2014.

[5] Hilary L Ashe, Joan Monks, Mark Wijgerde, Peter Fraser, and Nick J Proudfoot. Intergenic transcription and transinduction of the human β-globin locus. *Genes & development*, 11(19):2494–2509, October 1997.

[6] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue):W202–8, July 2009.

[7] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue):W369–73, July 2006.

[8] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 Dna-Sequences. *Cell*, 27(2):299–308, December 1981.

[9] Yoseph Barash, Gal Elidan, Nir Friedman, and Tommy Kaplan. *Modeling dependencies in protein-DNA binding sites*. ACM, New York, New York, USA, April 2003.

[10] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, and Ron Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research*, 37(Database issue):D885–D890, January 2009.

[11] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 1970.

[12] Emmanuel Beaudoing, Susan Freier, Jacqueline R Wyatt, Jean-Michel Claverie, and Daniel Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome research*, 10(7):1001–1010, July 2000.

[13] Michael G Berg, Larry N Singh, Ihab Younis, Qiang Liu, Anna Maria Pinto, Daisuke Kaida, Zhenxi Zhang, Sungchan Cho, Scott Sherrill-Mix, Lili Wan, and Gideon Dreyfuss. U1 snRNP Determines mRNA Length and Regulates Isoform Expression. *Cell*, 150(1):53–64, July 2012.

[14] O G Berg and P H von Hippel. Selection of DNA binding sites by regulatory proteins. *Trends in biochemical sciences*, 13(6):207–211, June 1988.

[15] Michael F Berger and Martha L Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols*, 4(3):393–411, 2009.

[16] Simon C Biddie, Sam John, Peter J Sabo, Robert E Thurman, Thomas A Johnson, R Louis Schiltz, Tina B Miranda, Myong-Hee Sung, Saskia Trump, Stafford L Lightman, Charles Vinson, John A Stamatoyannopoulos, and Gordon L Hager. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell*, 43(1):145–155, July 2011.

[17] Stefan Bonn, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczynski, Andrew Riddell, and Eileen E M Furlong. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156, February 2012.

[18] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*, 21(3):456–464, March 2011.

[19] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, September 2011.

[20] Eliezer Calo and Joanna Wysocka. Modification of Enhancer Chromatin: What, How, and Why? *Molecular cell*, 49(5):825–837, March 2013.

[21] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Par G Engstrom, Martin C Frith, Alistair R R Forrest, Wynand B Alkema, Sin L Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M Grimmond, Christine A Wells, Valerio Orlando, Claes Wahlestedt, Edison T Liu, Matthias Harbers, Jun Kawai, Vladimir B Bajic, David A Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635, June 2006.

[22] Rui Chen and Michael Snyder. Yeast proteomics and protein microarrays. *Journal of proteomics*, 73(11):2147–2157, October 2010.

[23] L Stirling Churchman and Jonathan S Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature Genetics*, 469(7330):368–373, January 2011.

[24] Thomas Clouaire, Shaun Webb, Pete Skene, Robert Illingworth, Alastair

Kerr, Robert Andrews, Jeong-Heon Lee, David Skalnik, and Adrian Bird. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes & development*, 26(15):1714–1728, August 2012.

[25] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature Genetics*, 489(7414):57–74, September 2012.

[26] Leighton J Core, Joshua J Waterfall, Daniel A Gilchrist, David C Fargo, Hojoong Kwak, Karen Adelman, and John T Lis. Defining the Status of RNA Polymerase at Promoters. *Cell Reports*, 2(4):1025–1035, October 2012.

[27] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, December 2008.

[28] Benoit Coulombre and Zachary F Burton. DNA bending and wrapping around RNA polymerase: A "revolutionary" model describing transcriptional mechanisms. *Microbiology and Molecular Biology Reviews*, 63(2):457–478, June 1999.

[29] Menno P Creyghton, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, Phillip A Sharp, Laurie A Boyer, Richard A Young, and Rudolf Jaenisch. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50):21931–21936, December 2010.

[30] Christian Kroun Damgaard, Søren Kahns, Søren Lykke-Andersen, Anders Lade Nielsen, Torben Heick Jensen, and Jørgen Kjems. A 5' Splice Site Enhances the Recruitment of Basal Transcription Initiation Factors In Vivo. *Molecular cell*, 29(2):271–278, February 2008.

[31] Charles G Danko, Stephanie Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Kevin Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. Accurate Identification of Active Transcriptional Regulatory Elements from Global Run-On Sequencing Data. *Manuscript in preparation*.

[32] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from

incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[33] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Roeder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Nadav S Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J Luo, Eddie Park, Kimberly Persaud, Jonathan B Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigo, and Thomas R Gingeras. Landscape of transcription in human cells. *Nature Genetics*, 489(7414):101–108, 2012.

[34] R Durbin, S Eddy, A Krogh, and G Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[35] Fred Ederer and Nathan Mantel. Confidence limits on the ratio of two Poisson variables. *American Journal of Epidemiology*, 100(3):165–167, 1974.

[36] Yasuaki Enoki and Hiroshi Sakurai. Diversity in DNA recognition by heat shock transcription factors (HSFs) from model organisms. *FEBS Letters*, 585(9):1293–1298, May 2011.

[37] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3):215–216, February 2012.

[38] Katalin Fejes-Toth, Vihra Sotirova, Ravi Sachidanandam, Gordon Assaf, Gregory J Hannon, Philipp Kapranov, Sylvain Foissac, Aarron T Willingham, Radha Duttagupta, Erica Dumais, and Thomas R Gingeras. Post-

transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature Genetics*, 457(7232):1028–1032, February 2009.

[39] Yair Field, Eilon Sharon, and Eran Segal. How Transcription Factors Identify Regulatory Sites in Genomic Sequence. In *A Handbook of Transcription Factors*, pages 193–204. Springer Netherlands, Dordrecht, January 2011.

[40] Peter C FitzGerald, David Sturgill, Andrey Shyakhtenko, Brian Oliver, and Charles Vinson. Comparative genomics of Drosophila and human core promoters. *Genome Biology*, 7(7):R53, July 2006.

[41] Joseph W Foley and Arend Sidow. Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC genomics*, 14(1):720, October 2013.

[42] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, January 2010.

[43] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, September 2008.

[44] M Fritsch and C Wu. Phosphorylation of Drosophila heat shock transcription factor. *Cell stress & chaperones*, 4(2):102–117, 1999.

[45] Nicholas J Fuda, M Behfar Ardehali, and John T Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature Genetics*, 461(7261):186–192, September 2009.

[46] Daniel J Gaffney, Graham McVicker, Athma A Pai, Yvonne N Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K Pritchard. Controls of Nucleosome Positioning in the Human Genome. *PLoS genetics*, 8(11):e1003036, November 2012.

[47] Daniel A Gilchrist, Gilberto Dos Santos, David C Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–551, November 2010.

[48] Sarah E Gonsalves, Alan M Moses, Zak Razak, Francois Robert, and J Timothy Westwood. Whole-Genome Analysis Reveals That Active Heat

Shock Factor Binding Sites Are Mostly Associated with Non-Heat Shock Genes in Drosophila melanogaster. *PLoS One*, 6(1):e15934, January 2011.

[49] Raluca Gordân, Alexander J Hartemink, and Martha L Bulyk. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome research*, 19(11):2090–2100, November 2009.

[50] Phil Green, Brent Ewing, Webb Miller, Pamela J Thomas, NISC Comparative Sequencing Program, and Eric D Green. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, 33(4):514–517, April 2003.

[51] Ulrike Grömping. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician*, 61(2):139–147, May 2007.

[52] Ana Rita Grosso, Sérgio Fernandes de Almeida, José Braga, and Maria Carmo-Fonseca. Dynamic transitions in RNA polymerase II density profiles during transcription termination. *Genome research*, 22(8):1447–1456, August 2012.

[53] Michael J Guertin and John T Lis. Chromatin Landscape Dictates HSF Binding to Target DNA Elements. *PLoS genetics*, 6(9):e1001114, September 2010.

[54] Michael J Guertin, S J Petesch, K L Zobeck, I M Min, and John T Lis. Drosophila Heat Shock System as a General Model to Investigate Transcriptional Regulation. *Cold Spring Harbor Symposia on Quantitative Biology*, 75(0):sqb.2010.75.039–9, April 2011.

[55] Nasun Hah, Charles G Danko, Leighton Core, Joshua J Waterfall, Adam Siepel, John T Lis, and W Lee Kraus. A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell*, 145(4):622–634, May 2011.

[56] Nasun Hah, Shino Murakami, Anusha Nagari, Charles G Danko, and W Lee Kraus. Enhancer transcripts mark active estrogen receptor binding sites. *Genome research*, 23(8):1210–1223, August 2013.

[57] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika

Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigo, and Tim J Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–1774, September 2012.

[58] Naoki Hayashida, Mitsuaki Fujimoto, Ke Tan, Ramachandran Prakasam, Toyohide Shinkawa, Liangping Li, Hitoshi Ichikawa, Ryosuke Takii, and Akira Nakai. Heat shock factor 1 ameliorates proteotoxicity in cooperation with the transcription factor NFAT. *The EMBO Journal*, 29(20):3459–3469, October 2010.

[59] Housheng H He, Clifford A Meyer, Hyunjin Shin, Shannon T Bailey, Gang Wei, Qianben Wang, Yong Zhang, Kexin Xu, Min Ni, Mathieu Lupien, Piotr Mieczkowski, Jason D Lieb, Keji Zhao, Myles Brown, and Xiaole S Liu. Nucleosome dynamics define transcriptional enhancers. *Nature Genetics*, 42:343–347, March 2010.

[60] Xin He, Chieh-Chun Chen, Feng Hong, Fang Fang, Saurabh Sinha, Huck-Hui Ng, and Sheng Zhong. A Biophysical Model for Analysis of Transcription Factor Interaction and Binding Site Arrangement from Genome-Wide Binding Data. *PLoS One*, 4(12):e8155, December 2009.

[61] Nathaniel D Heintzman and Bing Ren. Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*, 19(6):541–549, December 2009.

[62] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318, March 2007.

[63] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, Stanley Fields, and John A Stamatoyannopoulos. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Genetics*, 6(4):283–289, March 2009.

[64] Lin Himmelmann. R HMM. http://cran.r-project.org/web/packages/HMM/index.html, 2010. [Online; accessed 1-June-2014].

[65] Gangqing Hu, Dustin E Schones, Kairong Cui, River Ybarra, Daniel Northrup, Qingsong Tang, Luca Gattinoni, Nicholas P Restifo, Suming Huang, and Keji Zhao. Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome research*, 21(10):1650–1658, October 2011.

[66] Dean A Jackson, Francisco J Iborra, Erik M M Manders, and Peter R Cook. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. *Molecular Biology of the Cell*, 9(6):1523–1536, June 1998.

[67] Sam John, Peter J Sabo, Thomas A Johnson, Myong-Hee Sung, Simon C Biddie, Stafford L Lightman, Ty C Voss, Sean R Davis, Paul S Meltzer, John A Stamatoyannopoulos, and Gordon L Hager. Interaction of the glucocorticoid receptor with the chromatin landscape. *Molecular cell*, 29:611–624, March 2008.

[68] Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, March 2011.

[69] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, June 2007.

[70] Jeff W Johnson. A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1):1–19, 2000.

[71] Tamar Juven-Gershon, Jer-Yuan Hsu, Joshua W M Theisen, and James T Kadonaga. The RNA polymerase II core promoter—the gateway to transcription. *Current Opinion in Cell Biology*, 20(3):253–259, June 2008.

[72] Daisuke Kaida, Michael G Berg, Ihab Younis, Mumtaz Kasim, Larry N Singh, Lili Wan, and Gideon Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature Genetics*, 468(7324):664–U81, December 2010.

[73] Tommy Kaplan, Xiao-Yong Li, Peter J Sabo, Sean Thomas, John A Stamatoyannopoulos, Mark D Biggin, and Michael B Eisen. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS genetics*, 7(2):e1001290, February 2011.

[74] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, Ian Bell, Evelyn Cheung, Jorg Drenkow, Erica Dumais, Sandeep Patel, Gregg Helt, Madhavan Ganesh, Srinka Ghosh, Antonio Piccolboni, Victor Sementchenko, Hari Tammana, and Thomas R Gingeras. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488, June 2007.

[75] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome research*, (12):996–1006, 2002.

[76] W James Kent, A S Zweig, G Barber, A S Hinrichs, and D Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, September 2010.

[77] Peter V Kharchenko, Artyom A Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C Riddle, Jason Ernst, Peter J Sabo, Erica Larschan, Andrey A Gorchakov, Tingting Gu, Daniela Linder-Basso, Annette Plachetka, Gregory Shanower, Michael Y Tolstorukov, Lovelace J Luquette, Ruibin Xi, Youngsook L Jung, Richard W Park, Eric P Bishop, Theresa K Canfield, Richard Sandstrom, Robert E Thurman, David M MacAlpine, John A Stamatoyannopoulos, Manolis Kellis, Sarah C R Elgin, Mitzi I Kuroda, Vincenzo Pirrotta, Gary H Karpen, and Peter J Park. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature Genetics*, 471(7339):480–485, March 2011.

[78] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptewicz, Kellie Barbara-Haley, Scott Kuersten, Eirene Markenscoff-Papadimitriou, Dietmar Kuhl, Haruhiko Bito, Paul F Worley, Gabriel Kreiman, and Michael E Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature Genetics*, 465(7295):182–187, May 2010.

[79] Frederic Koch, Romain Fenouil, Marta Gut, Pierre Cauchy, Thomas K Albert, Joaquin Zacarias Cabeza, Salvatore Spicuglia, Albane Lamy de la Chapelle, Martin Heidemann, Corinna Hintermair, Dirk Eick, Ivo Gut,

Pierre Ferrier, and Jean Christophe Andrau. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural & Molecular Biology*, 18(8):956–U124, August 2011.

[80] William S Kruesi, Leighton J Core, Colin T Waters, John T Lis, and Barbara J Meyer. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife*, 2(0):–e00808, June 2013.

[81] Hojoong Kwak, Nicholas J Fuda, Leighton J Core, and John T Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122):950–953, February 2013.

[82] Hyun-sook Lee, Kevin W Kraus, Mariana F Wolfner, and John T Lis. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes & development*, 6(2):284–295, February 1992.

[83] Matthieu Legendre and Daniel Gautheret. Sequence determinants in human polyadenylation site selection. *BMC genomics*, 4(1):7, February 2003.

[84] Guoliang Li, Xiaoan Ruan, Raymond K Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, Yufen Goh, Joanne Lim, Jingyao Zhang, Hui Shan Sim, Su Qin Peh, Fabianus Hendriyan Mulawadi, Chin Thing Ong, Yuriy L Orlov, Shuzhen Hong, Zhizhuo Zhang, Steve Landt, Debasish Raha, Ghia Euskirchen, Chia-Lin Wei, Weihong Ge, Huaien Wang, Carrie Davis, Katherine I Fisher-Aylor, Ali Mortazavi, Mark Gerstein, Thomas Gingeras, Barbara Wold, Yi Sun, Melissa J Fullwood, Edwin Cheung, Edison Liu, Wing-Kin Sung, Michael Snyder, and Yijun Ruan. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, 148(1-2):84–98, 2012.

[85] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[86] Xiao-Yong Li, Sean Thomas, Peter J Sabo, Michael B Eisen, John A Stamatoyannopoulos, and Mark D Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology*, 12(R34):1–17, 2011.

[87] Syr-yaung Lin and Arthur D Riggs. The general affinity of *lac* repressor for E. coli DNA: Implications for gene regulation in procaryotes and eucaryotes. *Cell*, 4(2):107–111, February 1975.

[88] Yuefeng Lin, Zhihua Li, Fatih Ozsolak, Sang Woo Kim, Gustavo Arango-Argoty, Teresa T Liu, Scott A Tenenbaum, Timothy Bailey, A Paula Monaghan, Patrice M Milos, and Bino John. An in-depth map of polyadenylation sites in cancer. *Nucleic acids research*, 40(17):8460–8471, September 2012.

[89] Jiajian Liu and Gary D Stormo. Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic acids research*, 33(17):e141, 2005.

[90] Xiao Liu, David M Noll, Jason D Lieb, and Neil D Clarke. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome research*, 15(3):421–427, March 2005.

[91] Paul C Lott and Ian Korf. StochHMM: a flexible hidden Markov model tool and C++ Library. *Bioinformatics*, 30(11):btu057–1626, January 2014.

[92] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 resolution. *Nature Genetics*, 389(6648):251–260, September 1997.

[93] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1):10, August 2011.

[94] Kazuo Maruyama and Sumio Sugano. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene*, (138):171–174, 1994.

[95] Anthony Mathelier and Wyeth W Wasserman. The Next Generation of Transcription Factor Binding Site Prediction. *PLoS computational biology*, 9(9):e1003214, September 2013.

[96] Alexander Meissner, Tarjei S Mikkelsen, Hongcang Gu, Marius Wernig, Jacob Hanna, Andrey Sivachenko, Xiaolan Zhang, Bradley E Bernstein, Chad Nusbaum, David B Jaffe, Andreas Gnirke, Rudolf Jaenisch, and Eric S Lander. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature Genetics*, 454(7205):766–770, July 2008.

[97] Michael F Melgar, Francis S Collins, and Praveen Sethupathy. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biology*, 12(11):R113, November 2011.

[98] Jean-Francois Millau and Luc Gaudreau. CTCF, cohesin, and histone variants: connecting the genome. *Biochemistry and Cell Biology*, 89(5):505–513, October 2011.

[99] Leelavati Narlikar, Raluca Gordân, and Alexander J Hartemink. A Nucleosome-Guided Map of Transcription Factor Binding Sites in Yeast. *PLoS computational biology*, 3(11):e215, November 2007.

[100] Evgenia Ntini, Aino I Jaervelin, Jette Bornholdt, Yun Chen, Mette Boyd, Mette Jorgensen, Robin Andersson, Ilka Hoof, Aleks Schein, Peter R Andersen, Pia K Andersen, Pascal Preker, Eivind Valen, Xiaobei Zhao, Vicent Pelechano, Lars M Steinmetz, Albin Sandelin, and Torben Heick Jensen. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nature Structural & Molecular Biology*, 20(8):923–928, August 2013.

[101] Stuart H Orkin. Regulation of Globin Gene Expression in Erythroid Cells. *European Journal of Biochemistry*, 231(2):271–281, July 1995.

[102] Ulf Andersson Ørom, Thomas Derrien, Malte Beringer, Kiranmai Gumireddy, Alessandro Gardini, Giovanni Bussotti, Fan Lai, Matthias Zytnicki, Cedric Notredame, Qihong Huang, Roderic Guigo, and Ramin Shiekhattar. Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58, October 2010.

[103] Aleksandra Pekowska, Touati Benoukraf, Joaquin Zacarias Cabeza, Mohamed Belhocine, Frederic Koch, Hélène Holota, Jean Imbert, Jean Christophe Andrau, Pierre Ferrier, and Salvatore Spicuglia. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal*, 30(20):4198–4210, October 2011.

[104] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11S):S22–S32, November 2009.

[105] Jason Piper, Markus C Elze, Pierre Cauchy, Peter N Cockerill, Constanze Bonifer, and Sascha Ott. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, 41(21):e201, November 2013.

[106] Roger Pique-Regi, Jacob F Degner, Athma A Pai, Daniel J Gaffney, Yoav Gilad, and Jonathan K Pritchard. Accurate inference of transcription fac-

tor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3):447–455, March 2011.

[107] Andy Pohl and Miguel Beato. bwtool: a tool for bigWig files. *Bioinformatics*, 30(11):1618–1619, June 2014.

[108] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, January 2010.

[109] Pascal Preker, Jesper Nielsen, Susanne Kammler, Søren Lykke-Andersen, Marianne S Christensen, Christophe K Mapendano, Mikkel H Schierup, and Torben Heick Jensen. RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909):1851–1854, December 2008.

[110] Lawrence Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[111] Alvaro Rada-Iglesias, Ruchi Bajpai, Tomek Swigut, Samantha A Brugmann, Ryan A Flynn, and Joanna Wysocka. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature Genetics*, 470(7333):279–283, February 2011.

[112] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA Synthesis in Mammalian Cells. *PLoS biology*, 4(10):e309, September 2006.

[113] Robin Reed. Coupling transcription, splicing and mRNA export. *Current Opinion in Cell Biology*, 15(3):326–331, June 2003.

[114] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. 147(6):1408–1419, December 2011.

[115] Patricia Richard and James L Manley. Transcription termination by nuclear RNA polymerases. *Genes & development*, 23(11):1247–1269, June 2009.

[116] Adam Roberts, Harold Pimentel, Cole Trapnell, and Lior Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, September 2011.

[117] A Gordon Robertson, Mikhail Bilenky, Angela Tam, Yongjun Zhao, Thomas Zeng, Nina Thiessen, Timothee Cezard, Anthony P Fejes, Elizabeth D Wederell, Rebecca Cullum, Ghia Euskirchen, Martin Krzywinski, Inanc Birol, Michael Snyder, Pamela A Hoodless, Martin Hirst, Marco A Marra, and Steven J M Jones. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome research*, 18(12):1906–1917, December 2008.

[118] Emmanuelle Roulet, Stéphane Busso, Anamaria A Camargo, Andrew J G Simpson, Nicolas Mermod, and Philipp Bucher. High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, 20(8):831–834, August 2002.

[119] M K Sakharkar, VTK Chow, and P Kangueane. Distributions of exons and introns in the human genome. *In silico biology*, 4(4):387–393, 2004.

[120] Hiroshi Sakurai and Yukiko Takemori. Interaction between heat shock transcription factors (HSFs) and divergent binding sequences: binding specificities of yeast HSFs and human HSF1. *The Journal of biological chemistry*, 282(18):13334–13341, May 2007.

[121] Albin Sandelin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David A Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature Reviews Genetics*, 8(6):424–436, June 2007.

[122] Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, March 2008.

[123] Amy C Seila, J Mauro Calabrese, Stuart S Levine, Gene W Yeo, Peter B Rahl, Ryan A Flynn, Richard A Young, and Phillip A Sharp. Divergent transcription from active promoters. *Science*, 322(5909):1849–1851, December 2008.

[124] Eilon Sharon, Shai Lubliner, and Eran Segal. A Feature-Based Approach to Modeling Protein–DNA Interactions. *PLoS computational biology*, 4(8):e1000154, August 2008.

[125] Ali Shilatifard. The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. *Annual Review of Biochemistry, Vol 81*, 81(1):65–95, 2012.

[126] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15776–15781, December 2003.

[127] Daria Shlyueva, Gerald Stampfel, and Alexander Stark. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286, April 2014.

[128] Michaela Smolle and Jerry L Workman. Transcription-associated histone modifications and cryptic transcription. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(1):84–97, January 2013.

[129] Lincoln D Stein. Bio-BigFile. `http://search.cpan.org/~lds/Bio-BigFile-1.01/`, 2010. [Online; accessed 1-June-2014].

[130] Andrew B Stergachis, Shane Neph, Alex Reynolds, Richard Humbert, Brady Miller, Sharon L Paige, Benjamin Vernot, Jeffrey B Cheng, Robert E Thurman, Richard Sandstrom, Eric Haugen, Shelly Heimfeld, Charles E Murry, Joshua M Akey, and John A Stamatoyannopoulos. Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. 154(4):888–903, August 2013.

[131] Regina Stoltenburg, Christine Reinemann, and Beate Strehlitz. SELEX—a (r) evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular engineering*, 24(4):381–403, October 2007.

[132] Gary D Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000.

[133] James Taylor, Bob Harris, David King, Brent Pedersen, and et. al Li, Kanwei. bx-python. `https://pypi.python.org/pypi/bx-python/0.7.1`, 2011. [Online; accessed 1-June-2014].

[134] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Jornal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[135] Scott Tonidandel and James M LeBreton. Relative Importance Analysis:

A Useful Supplement to Regression Analysis. *Journal of Business and Psychology*, 26(1):1–9, March 2011.

[136] Baris Tursun, Tulsi Patel, Paschalis Kratsios, and Oliver Hobert. Direct conversion of C. elegans germ cells into specific neuron types. *Science*, 331(6015):304–308, January 2011.

[137] Eivind Valen, Albin Sandelin, Ole Winther, and Anders Krogh. Discovery of Regulatory Elements is Improved by a Discriminatory Approach. *PLoS computational biology*, 5(11):e1000562, November 2009.

[138] Bryan J Venters and B Franklin Pugh. Genomic organization of human transcription initiation complexes. *Nature Genetics*, 502(7469):53–58, October 2013.

[139] A J Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, April 1967.

[140] P H von Hippel, A Revzin, and C A Gross. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: 1. The lac operon: equilibrium aspects. In *Proceedings of the Nacional Academy of Science*, pages 4808–4812, December 1974.

[141] Ty C Voss, R Louis Schiltz, Myong-Hee Sung, Paul M Yen, John A Stamatoyannopoulos, Simon C Biddie, Thomas A Johnson, Tina B Miranda, Sam John, and Gordon L Hager. Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism. *Cell*, 146(4):544–554, August 2011.

[142] Joseph A Wamstad, Jeffrey M Alexander, Rebecca M Truty, Avanti Shrikumar, Fugen Li, Kirsten E Eilertson, Huiming Ding, John N Wylie, Alexander R Pico, John A Capra, Genevieve Erwin, Steven J Kattman, Gordon M Keller, Deepak Srivastava, Stuart S Levine, Katherine S Pollard, Alisha K Holloway, Laurie A Boyer, and Benoit G Bruneau. Dynamic and Coordinated Epigenetic Regulation of Developmental Transitions in the Cardiac Lineage. *Cell*, 151(1):206–220, September 2012.

[143] Dong Wang, Ivan Garcia-Bassets, Chris Benner, Wenbo Li, Xue Su, Yiming Zhou, Jinsong Qiu, Wen Liu, Minna U Kaikkonen, Kenneth A Ohgi, Christopher K Glass, Michael G Rosenfeld, and Xiang-Dong Fu. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature Genetics*, 474(7351):390–394, May 2011.

[144] Isabel X Wang, Leighton J Core, Hojoong Kwak, Lauren Brady, Alan Bruzel, Lee McDaniel, Allison L Richards, Ming Wu, Christopher Grunseich, John T Lis, and Vivian G Cheung. RNA-DNA Differences Are Generated in Human Cells within Seconds after RNA Exits Polymerase II. *Cell Reports*, 6(5):906–915, March 2014.

[145] Weisheng Wu, Yong Cheng, Cheryl A Keller, Jason Ernst, Swathi Ashok Kumar, Tejaswini Mishra, Christapher Morrissey, Christine M Dorman, Kuan-Bei Chen, Daniela Drautz, Belinda Giardine, Yoichiro Shibata, Lingyun Song, Max Pimkin, Gregory E Crawford, Terrence S Furey, Manolis Kellis, Webb Miller, James Taylor, Stephan C Schuster, Yu Zhang, Francesca Chiaromonte, Gerd A Blobel, Mitchell J Weiss, and Ross C Hardison. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome research*, 21(10):1659–1671, October 2011.

[146] Xuebing Wu and Phillip A Sharp. Divergent Transcription: A Driving Force for New Gene Origination? *Cell*, 155(5):990–996, 2013.

[147] Gabriel E Zentner, Paul J Tesar, and Peter C Scacheri. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–1283, August 2011.

[148] Daniel R Zerbino, Nathan Johnson, Thomas Juettemann, Steven P Wilder, and Paul Flicek. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*, 30(7):btt737–1009, December 2013.

[149] Yong Zhang, Tao Liu, Clifford A Meyer, Jérome Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, September 2008.

[150] Qing Zhou and Wing H Wong. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 101(33):12114–12119, August 2004.